

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されて
いる事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed
with this Office.

出願年月日 2002年12月17日
Date of Application:

出願番号 特願2002-365764
Application Number:

パリ条約による外国への出願
に用いる優先権の主張の基礎
となる出願の国コードと出願
番号
The country code and number
of your priority application,
to be used for filing abroad
under the Paris Convention, is

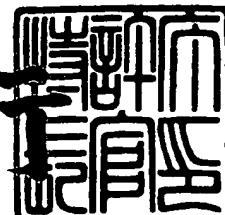
J P 2 0 0 2 - 3 6 5 7 6 4

出願人 独立行政法人科学技術振興機構
Applicant(s):

2011年 4月12日

特許庁長官
Commissioner,
Japan Patent Office

岩井良行



出証番号 出証特2011-3012903

【書類名】 特許願
【整理番号】 P2234JST
【特記事項】 特許法第30条第1項の規定の適用を受けようとする特許出願
【提出日】 平成14年12月17日
【あて先】 特許庁長官 殿
【発明者】
 【住所又は居所】 千葉県佐倉市臼井86
 【氏名】 中臺 一博
【発明者】
 【住所又は居所】 東京都渋谷区西原2-10-9
 【氏名】 奥乃 博
【発明者】
 【住所又は居所】 埼玉県川越市西小仙波町2-18-3
 【氏名】 北野 宏明
【特許出願人】
 【識別番号】 396020800
 【氏名又は名称】 科学技術振興事業団
【代理人】
 【識別番号】 100082876
 【弁理士】
 【氏名又は名称】 平山 一幸
 【電話番号】 03-3352-1808
【選任した代理人】
 【識別番号】 100069958
 【弁理士】
 【氏名又は名称】 海津 保三

【手数料の表示】**【予納台帳番号】** 031727**【納付金額】** 21,000円**【提出物件の目録】****【物件名】** 明細書 1**【物件名】** 図面 1**【物件名】** 要約書 1**【包括委任状番号】** 0013677**【ブルーフの要否】** 要

【書類名】 明細書

【発明の名称】 ロボット視聴覚システム

【特許請求の範囲】

【請求項1】 各話者が発した単語とその方向とを組み合わせて成る複数の音響モデルと、これらの音響モデルを使用して、音源分離された音響信号に対して音声認識プロセスを実行する音声認識エンジンと、前記音声認識プロセスによって前記音響モデル別に得られた複数の音声認識プロセス結果を統合し、何れかの音声認識プロセス結果を選択するセレクタと、を備え、

各話者が同時に発話した単語を各々認識することを特徴とする、ロボット視聴覚システム。

【請求項2】 前記セレクタが、多数決により前記音声認識プロセス結果を選択するように構成されていることを特徴とする、請求項1に記載のロボット視聴覚システム。

【請求項3】 前記セレクタにて選択された音声認識プロセス結果を外部に出力する対話部を備えていることを特徴とする、請求項1又は2に記載のロボット視聴覚システム。

【請求項4】 外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいて、ピッチ抽出、調波構造に基づいたグループングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定し、その聴覚イベントを抽出する聴覚モジュールと、

ロボットの前方を撮像するカメラを備えており、このカメラにより撮像された画像に基づいて各話者の顔識別と定位とから、各話者を同定してその顔イベントを抽出する顔モジュールと、

ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモータイベントを抽出するモータ制御モジュールと、

上記聴覚イベント、顔イベント及びモータイベントから、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定し、この決定に対してカルマンフィルタを用いて上記イベントを時間方向に接続することにより聴覚ストリーム及び顔ストリームを生成し、さらにこれらを関連付けてア

ソシエーションストリームを生成するアソシエーションモジュールと、これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいて、モータの駆動制御を行うアテンション制御モジュールと、を備え、

上記聴覚モジュールが、

上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（IPD）または両耳間強度差（IID）をもったサブバンドを集めて、音源の波形を再構築することにより、音源分離を行なうと共に、

複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルによる音声認識結果をセレクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されていることを特徴とする、ロボット視聴覚システム。

【請求項5】 外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいて、ピッチ抽出、調波構造に基づいたグループングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定し、その聴覚イベントを抽出する聴覚モジュールと、

ロボットの前方を撮像するカメラを備えており、このカメラにより撮像された画像に基づいて各話者の顔識別と定位とから、各話者を同定してその顔イベントを抽出する顔モジュールと、

ステレオカメラにより撮像された画像から抽出された視差に基づいて縦に長い物体を抽出定位して、ステレオイベントを抽出するステレオモジュールと、

ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモータイベントを抽出するモータ制御モジュールと、

上記聴覚イベント、顔イベント、ステレオイベント及びモータイベントから、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定し、この決定に対してカルマンフィルタを用いて上記イベントを時間方向に接続することにより聴覚ストリーム、顔ストリーム及びステレオ視覚スト

リームを生成し、さらにこれらを関連付けてアソシエーションストリームを生成するアソシエーションモジュールと、

これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいて、モータの駆動制御を行うアテンション制御モジュールと、を備え、

上記聴覚モジュールが、

上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（IPD）または両耳間強度差（IID）をもったサブバンドを集めて、音源の波形を再構築することにより音源分離を行なうと共に、

複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルによる音声認識結果をセレクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されていることを特徴とする、ロボット視聴覚システム。

【請求項6】 前記聴覚モジュールによる音声認識ができなかったときに、前記アテンション制御モジュールが、当該音響信号の音源の方向に前記マイク及び前記カメラを向けて前記マイクから再び音声を集めさせ、この音に対して前記聴覚モジュールにより音源定位・分離された音響信号に基づいて再度聴覚モジュールによる音声認識を行なうように構成されていることを特徴とする、請求項4又は5に記載のロボット視聴覚システム。

【請求項7】 前記聴覚モジュールが、音声認識を行なう際に、顔モジュールによる顔イベント又は／及びステレオモジュールによるステレオイベントを参照することを特徴とする、請求項5又は6に記載のロボット視聴覚システム。

【請求項8】 前記聴覚モジュールにて判断された音声認識結果を外部に出力する対話部を備えていることを特徴とする、請求項4～7の何れかに記載のロボット視聴覚システム。

【請求項9】 前記アクティブ方向通過型フィルタのパスレンジが、周波数毎に制御可能であることを特徴とする、請求項4～8の何れかに記載のロボット

視聴覚システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明はロボット、特に人型または動物型ロボットにおける視聴覚システムに関するものである。

【0002】

【従来の技術】

近年、このような人型または動物型ロボットにおいては、AIの研究目的の対象にとどまらず、所謂「人間のパートナー」としての将来的な利用が考えられている。そして、ロボットが人間との知的なソーシャルインタラクションを行なうために、視聴覚等の知覚がロボットには必要である。そして、ロボットが人間とのソーシャルインタラクションを実現するためには、知覚のうち、視聴覚、特に聴覚が重要な機能であることは明らかである。従って、視覚、聴覚に関して、所謂能動知覚が注目されてきている。

【0003】

ここで、能動知覚とは、ロボット視覚やロボット聴覚等の知覚を担当する知覚装置を知覚すべき目標に追従する働きをを言い、例えば、これらの知覚装置を支持する頭部を駆動機構により目標に追従するように姿勢制御するものである。

【0004】

ロボットにおける能動視覚においては、少なくとも知覚装置であるカメラが、駆動機構による姿勢制御によって、その光軸方向を目標に向かって保持され、更に目標に対して自動的にフォーカシングやズームイン、ズームアウト等を行う。これにより、目標が移動してもカメラによって撮像される。このような能動視覚の研究が従来、様々に行なわれている。

【0005】

これに対して、ロボットにおける能動聴覚においては、少なくとも知覚装置であるマイクが、駆動機構による姿勢制御によってその指向性を目標に向かって保持され、目標からの音がマイクによって集音される。このとき、能動聴覚の不利

な点として、駆動機構が作用している間はマイクが駆動機構の作動音を拾ってしまうために目標からの音に比較的大きなノイズが混入してしまい、目標からの音を認識できなくなってしまうことがある。このような能動聴覚の不利な点を排除するために、例えば視覚情報を参照して音源の方向付けを行なうことにより、目標からの音を正確に認識する方法が採用されている。

【0006】

ところで、このような能動聴覚においては、マイクで集音した音に基づいて、

(A) 音源の定位, (B) 各音源から発せられた音毎の分離, (C) そして各音源からの音の認識を行なう必要がある。このうち、(A) 音源定位及び(B) 音源分離については、能動聴覚における実時間・実環境での音源定位・追跡・分離に関する種々の研究が行なわれている（特許文献1参照）。

【0007】

【特許文献1】

国際公開第01/95314号パンフレット

【0008】

ここで、例えば、特許文献1に示すように、HRTF（頭部伝達関数）から求められる両耳間位相差（IPD）、両耳間強度差（IID）を利用して音源定位を行なうことが知られている。また、特許文献1では、例えば所謂方向通過型フィルタ、即ちディレクションパスフィルタを用いて、特定の方向のIPDと同じIPDを有するサブバンドを選択することにより、各音源からの音を分離する方法が知られている。

【0009】

これに対して、音源分離により分離された各音源からの音の認識については、例えばマルチコンディショニングやミッシングデータ等のノイズに対してロバストな音声認識へのアプローチは種々の研究が行なわれている（例えば非特許文献1, 2参照）。

【0010】

【非特許文献1】

J. ベーカー等著、クリーンスピーチモデルに基づくロバスト“ユー

ロスピーチ2001—第7回ヨーロッパ会議予稿集”，2001年，第1巻，p 213—216 (J.Baker, M.Cooke, and P.Green, Robust as based on clean speechmodels: An evaluation of missing data techniques for connected digit recognition in noise. "7th European conference on Speech Communication Technology", Volume 1, p. 213-216)

【非特許文献2】

P. レネベイ等著，ロバストスピーチ認識”ユーロスピーチ2001—第7回ヨーロッパ会議予稿集”，2001年，第12巻，pp. 1107—1110 (Philippe Renevey, Rolf Vetter, and Jens Kraus. Robust speech recognition using missing feature theory and vector quantization. "7th European Conference on Speech Communication Technology", Volume 12, pp. 1107-1110)

【0011】

【発明が解決しようとする課題】

しかしながら、これらの研究（例えば非特許文献1，2）においては、S／N比が小さい場合には、有効な音声認識を行なうことができない。また、実時間・実環境での音声認識についての研究は行なわれていない。

【0012】

この発明は、以上の点に鑑みて、各音源からの分離された音についての認識を行なうようにしたロボット視聴覚システムを提供することを目的としている。

【0013】

【課題を解決するための手段】

上記目的を達成するために、本発明のロボット視聴覚システムの第1の構成は、各話者が発した単語とその方向とを組み合わせて成る複数の音響モデルと、これらの音響モデルを使用して、音源分離された音響信号に対して音声認識プロセスを実行する音声認識エンジンと、この音声認識プロセスによって音響モデル別に得られた複数の音声認識プロセス結果を統合し、何れかの音声認識プロセス結果を選択するセレクタと、を備えて、各話者が同時に発話した単語を各々認識することを特徴としている。

【0014】

前記セレクタは、多数決により前記音声認識プロセス結果を選択するように構成され、前記セレクタにて選択された音声認識プロセス結果を外部に出力する対話部を備えていてもよい。

【0015】

このような第1の構成によれば、音源定位・音源分離された音響信号に基づいて、複数の音響モデルを使用することによって、それぞれ音声認識プロセスを行なう。そして、各音響モデルによる音声認識プロセス結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断する。

【0016】

また、上記目的を達成するために、本発明のロボット視聴覚システムの第2の構成は、外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいて、ピッチ抽出、調波構造に基づいたグルーピングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定し、その聴覚イベントを抽出する聴覚モジュールと、ロボットの前方を撮像するカメラを備えており、このカメラにより撮像された画像に基づいて各話者の顔識別と定位とから各話者を同定してその顔イベントを抽出する顔モジュールと、ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモータイベントを抽出するモータ制御モジュールと、上記聴覚イベント、顔イベント及びモータイベントから、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定し、この決定に対してカルマンフィルタを用いて上記イベントを時間方向に接続することにより聴覚ストリーム及び顔ストリームを生成し、さらにこれらを関連付けてアソシエーションストリームを生成するアソシエーションモジュールと、これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいてモータの駆動制御を行うアテンション制御モジュールと、を備え、上記聴覚モジュールが、上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（IPD）または

両耳間強度差（I I D）をもったサブバンドを集めて、音源の波形を再構築することにより音源分離を行なうと共に、複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルによる音声認識結果をセレクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されている。

【0017】

このような第2の構成によれば、聴覚モジュールがマイクが集音した外部の対象からの音から、調波構造を利用してピッチ抽出を行なうことにより音源毎の方向を得て個々の話者を同定し、その聴覚イベントを抽出する。

【0018】

また、顔モジュールが、カメラにより撮像された画像から、パターン認識による各話者の顔識別と定位から、個々の話者の顔イベントを抽出する。

【0019】

さらに、モータ制御モジュールが、ロボットを水平方向に回動させる駆動モータの回転位置に基づいて、ロボットの方向を検出することによって、モータイベントを抽出する。

【0020】

なお、上記イベントとは、各時点において検出される音または顔が在ること、あるいは駆動モータが回転される状態を示しており、ストリームとは、エラー訂正処理を行ないながら例えばカルマンフィルタ等により時間的に連続するよう接続したイベントを示している。

【0021】

ここで、アソシエーションモジュールは、このようにしてそれぞれ抽出された聴覚イベント、顔イベント及びモータイベントに基づいて、各話者の聴覚ストリーム及び顔ストリームを生成し、さらにこれらのストリームを関連付けてアソシエーションストリームを生成して、アテンション制御モジュールがこれらのストリームに基づいてアテンション制御を行なうことにより、モータ制御モジュールの駆動モータ制御のプランニングを行なう。なお、アソシエーションストリームとは、聴覚ストリーム及び顔ストリームを包含する概念である。

【0022】

なお、アテンションとは、ロボットが対象である話者を聴覚的及び／または視覚的に「注目」することであり、アンテンション制御とは、モータ制御モジュールによりその向きを変えることによってロボットが上記話者に注目するようになることである。

【0023】

そして、アテンション制御モジュールは、このプランニングに基づいて、モータ制御モジュールの駆動モータを制御することにより、ロボットの方向を対象である話者に向ける。これにより、ロボットが対象である話者に対して正対することにより、聴覚モジュールが当該話者の声を感度の高い正面方向にてマイクで正確に集音、定位することができると共に、顔モジュールが当該話者の画像をカメラにより良好に撮像することができるようになる。

【0024】

従って、このような聴覚モジュール、顔モジュール及びモータ制御モジュールと、アソシエーションモジュール及びアテンション制御モジュールとの連携によって、ロボットの聴覚及び視覚がそれぞれ有する曖昧性が互いに補完されることになり、所謂ロバスト性が向上し、複数の話者であっても、各話者をそれぞれ知覚することができる。

【0025】

また、例えば聴覚イベントまたは顔イベントの何れか一方が欠落したときであっても、顔イベントまたは聴覚イベントのみに基づいて、対象である話者をアソシエーションモジュールが知覚することができるので、リアルタイムにモータ制御モジュールの制御を行なうことができる。

【0026】

さらに、上記聴覚モジュールが、上述したように音源定位・音源分離された音響信号に基づいて、複数の音響モデルを使用することによってそれぞれ音声認識を行なう。そして、各音響モデルによる音声認識結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断する。

【0027】

これにより、従来の音声認識と比較して複数の音響モデルを使用することによって、実時間・実環境での正確な音声認識を行なうことが可能になると共に、各音響モデルによる音声認識結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断して、より一層正確な音声認識を行なうことができる。

【0028】

また、上記目的を達成するために、本発明のロボット視聴覚システムの第3の構成は、外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいてピッチ抽出、調波構造に基づいたグルーピングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定しその聴覚イベントを抽出する聴覚モジュールと、ロボットの前方を撮像するカメラを備えこのカメラで撮像された画像に基づいて各話者の顔識別と定位とから各話者を同定してその顔イベントを抽出する顔モジュールと、ステレオカメラにより撮像された画像から抽出された視差に基づいて縦に長い物体を抽出定位してステレオイベントを抽出するステレオモジュールと、ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモータイベントを抽出するモータ制御モジュールと、前記聴覚イベント、顔イベント、ステレオイベント及びモータイベントから聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定しこの決定に対してカルマンフィルタを用いて前記イベントを時間方向に接続することにより聴覚ストリーム、顔ストリーム及びステレオ視覚ストリームを生成しさらにこれらを関連付けてアソシエーションストリームを生成するアソシエーションモジュールと、これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいてモータの駆動制御を行うアテンション制御モジュールと、を備え、上記聴覚モジュールが、上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（I P D）または両耳間強度差（I I D）をもったサブバンドを集めて、音源の波形を再構築することにより音源分離を行なうと共に、音声認識の際に、複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルに

よる音声認識結果をセレクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されている。

【0029】

このような第3の構成によれば、聴覚モジュールは、マイクが集音した外部の目標からの音から調波構造を利用してピッチ抽出を行なうことにより音源毎の方向を得て、個々の話者的方向を決定してその聴覚イベントを抽出する。

【0030】

また、顔モジュールは、カメラにより撮像された画像からパターン認識による各話者の顔識別と定位から各話者を同定して、個々の話者の顔イベントを抽出する。さらに、ステレオモジュールは、ステレオカメラにより撮像された画像から抽出された視差に基づいて縦に長い物体を抽出定位してステレオイベントを抽出する。

【0031】

さらに、モータ制御モジュールは、ロボットを水平方向に回動させる駆動モータの回転位置に基づいて、ロボットの方向を検出することによってモータイベントを抽出する。

【0032】

なお、上記イベントとは、各時点において検出される音、顔及び縦に長い物体が在ること、あるいは駆動モータが回転される状態を示しており、ストリームとは、エラー訂正処理を行ないながら例えばカルマンフィルタ等により時間的に連続するように接続したイベントを示している。

【0033】

ここで、アソシエーションモジュールは、このようにしてそれぞれ抽出された聴覚イベント、顔イベント、ステレオイベント及びモータイベントに基づいて、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報によって各話者の方向を決定することにより、各話者の聴覚ストリーム、顔ストリーム及びステレオ視覚ストリームを生成し、さらにこれらのストリームを関連付けてアソシエーションストリームを生成する。なお、アソシエーションストリームとは、聴覚ストリーム、顔ストリーム及びステレオ視覚ストリームを包含する概念である。この

際、アソシエーションモジュールは、聴覚イベントの音源定位及び顔イベントの顔定位、即ち聴覚及び視覚の方向情報に基づいて各話者の方向を決定し、決定された各話者の方向を参考にして、アソシエーションストリームを生成する。

【0034】

そして、アテンション制御モジュールが、これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいて、モータの駆動制御を行なう。そして、アテンション制御モジュールは、このプランニングに基づいてモータ制御モジュールの駆動モータを制御してロボットの方向を目標である話者に向ける。これにより、ロボットが目標である話者に対して正対することによって聴覚モジュールが当該話者の声を感度の高い正面方向にてマイクにより正確に集音、定位することができる共に、顔モジュールが当該話者の画像をカメラにより良好に撮像することができるようになる。

【0035】

従って、このような聴覚モジュール、顔モジュール、ステレオモジュール及びモータ制御モジュールと、アソシエーションモジュール及びアテンション制御モジュールとの連携によって、聴覚ストリームの音源定位及び顔ストリームの話者定位という方向情報に基づいて各話者の方向を決定することにより、ロボットの聴覚及び視覚がそれぞれ有する曖昧性が互いに補完されることになり、所謂ロバスト性が向上し、複数の話者であっても各話者をそれぞれ確実に知覚することができる。

【0036】

また、例えば聴覚ストリーム、顔ストリーム及びステレオ視覚ストリームの何れかが欠落したときであっても、残りのストリームに基づいて目標である話者をアテンション制御モジュールが追跡することができるので、正確に目標の方向を把握して、モータ制御モジュールの制御を行なうことができる。

【0037】

ここで、聴覚モジュールが、アソシエーションモジュールからのアソシエーションストリームを参照することにより、顔モジュールからの顔ストリームやステレオモジュールからのステレオ視覚ストリームをも考慮して音源定位を行なうこと

とによって、より一層正確な音源定位を行なうことができる。

【0038】

そして、上記聴覚モジュールは、アソシエーションモジュールからの正確な音源方向情報に基づいて、聴覚特性に従って正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（IPD）または両耳間強度差（IID）をもったサブバンドを集めて、音源の波形を再構築して音源分離を行なうので、上述した聴覚特性に応じてパスレンジ即ち感度を調整することにより、方向による感度の違いを考慮して、より正確に音源分離を行なうことができる。さらに、上記聴覚モジュールは、上述したように聴覚モジュールによって音源定位・音源分離された音響信号に基づいて、複数の音響モデルを使用することによってそれぞれ音声認識を行なう。そして、各音響モデルによる音声認識結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断して、この音声認識結果を対応する話者と関連付けて出力する。

【0039】

これにより、従来の音声認識と比較して、複数の音響モデルを使用することによって、実時間・実環境での正確な音声認識を行なうことが可能になると共に、各音響モデルによる音声認識結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断することにより、より一層正確な音声認識を行なうことができる。

【0040】

なお、第2の構成と第3の構成においては、聴覚モジュールによる音声認識ができなかったときに、前記アテンション制御モジュールが、当該音響信号の音源の方向に前記マイク及び前記カメラを向けて、前記マイクから再び音声を集音させ、この音に対して聴覚モジュールにより音源定位・分離された音響信号に基づいて、再度聴覚モジュールによる音声認識を行なうように構成されている。

【0041】

さらに、前記聴覚モジュールは、音声認識を行なう際に顔モジュールによる顔イベントを参照するのが望ましい。また、前記聴覚モジュールにて判断された音

声認識結果を外部に出力する対話部が備えられていてもよい。さらに、前記アクティブ方向通過型フィルタのパスレンジが周波数毎に制御可能であることが望ましい。

【0042】

上記聴覚モジュールによる音声認識ができなかったとき、アテンション制御モジュールが、当該音響信号の音源の方向（当該話者）にマイク及びカメラを向けて、再度マイクから音声を集音させ、聴覚モジュールにより音源定位・分離された音響信号に基づいて、再度聴覚モジュールによる音声認識を行なう場合には、ロボットの聴覚モジュールのマイク及び顔モジュールのカメラが当該話者と正対することによって、確実な音声認識を行なうことが可能になる。

【0043】

上記聴覚モジュールは、音声認識を行なう際に、アソシエーションモジュールからのアソシエーションストリームを参照することにより、顔モジュールからの顔ストリームをも考慮する。即ち、聴覚モジュールは、顔モジュールにより定位された顔イベントに関して、聴覚モジュールにより定位・分離された音源（話者）からの音響信号に基づいて音声認識を行なうことにより、より一層正確な音声認識を行なうことができる。

【0044】

上記アクティブ方向通過型フィルタのパスレンジが周波数毎に制御可能であると、さらに集音した音からの分離の精度が上がり、これにより音声認識もさらに向上する。

【0045】

【発明の実施の形態】

以下、図面に示した実施形態に基づいて、この発明を詳細に説明する。

図1及び図2は、それぞれこの発明によるロボット視聴覚システムの一実施形態を備えた実験用の上半身のみの人型ロボットの全体構成例を示している。図1において、人型ロボット10は、4DOF（自由度）のロボットとして構成されており、ベース11と、ベース11上にて一軸（垂直軸）周りに回動可能に支持された胴体部12と、胴体部12上にて三軸方向（垂直軸、左右方向の水平軸及

び前後方向の水平軸）の周りに揺動可能に支持された頭部13とを含んでいる。

【0046】

上記ベース11は固定配置されていてもよく、脚部として動作可能としてもよい。また、ベース11は、移動可能な台車等の上に載置されていてもよい。胴体部12は、ベース11に対して垂直軸の周りに、図1にて矢印Aで示すように回動可能に支持されており、図示しない駆動手段によって回転駆動されると共に、図示の場合、防音性の外装によって覆われている。

【0047】

頭部13は胴体部12に対して連結部材13aを介して支持されており、この連結部材13aに対して前後方向の水平軸の周りに、図1にて矢印Bで示すように揺動可能に、また左右方向の水平軸の周りに、図2にて矢印Cで示すように揺動可能に支持されていると共に、上記連結部材13aが、胴体部12に対してさらに前後方向の水平軸の周りに、図1にて矢印Dで示すように揺動可能に支持されており、それぞれ図示しない駆動手段によって、各矢印A、B、C、D方向に回転駆動される。ここで、頭部13は、図3に示すように全体が防音性の外装14により覆われ、前側にロボット視覚を担当する視覚装置としてのカメラ15、両側にロボット聴覚を担当する聴覚装置としての一対のマイク16（16a、16b）を備えている。なお、マイク16は、頭部13の両側に限定されることなく、頭部13の他の位置あるいは胴体部12等に設けられていてもよい。

【0048】

上記外装14は、例えばウレタン樹脂等の吸音性の合成樹脂から構成されており、頭部13の内部がほぼ完全に密閉されて、頭部13の内部の遮音が行われるように構成されている。なお、胴体部12の外装も、頭部13の外装14と同様に、吸音性の合成樹脂から構成されている。

【0049】

上記カメラ15は公知の構成であって、例えば所謂パン、チルト、ズームの3DOF（自由度）を有する市販のカメラにより構成されている。尚、カメラ15は、同期をとってステレオ画像を送ることができるように設計されている。

【0050】

上記マイク16は、それぞれ頭部13の側面において前方に向かって指向性を有するように取り付けられている。マイク16の左右の各マイク16a, 16bは、それぞれ図1及び図2に示すように、頭部13の外装14の両側に配置された段部14a, 14bの内側に取り付けられている。そして、各マイク16a, 16bは、段部14a, 14bに設けられた貫通穴を通して、前方の音を集音すると共に、外装14の内部の音を拾わないように、適宜の手段により遮音されている。なお、段部14a, 14bに設けられた貫通穴は、段部14a, 14bの内側から頭部前方に向けて貫通するように、各段部14a, 14bに形成されている。これにより、各マイク16a, 16bは、所謂バイノーラルマイクとして構成されている。なお、マイク16a, 16bの取付位置に近接する外装14は人間の外耳形状に形成されていてもよい。ここで、マイク16は、外装14の内側に配置された一対の内部マイクを含んでいてもよく、この内部マイクにより集音された内部音に基づいて、ロボット10の内部に発生するノイズをキャンセルすることができる。

【0051】

図4は、上記カメラ15及びマイク16を含むロボット視聴覚の電気的構成例を示している。図4において、ロボット視聴覚システム17は、聴覚モジュール20, 顔モジュール30, ステレオモジュール37, モータ制御モジュール40及びアソシエーションモジュール50から構成されている。

【0052】

ここで、アソシエーションモジュール50はクライアントからの依頼に応じて処理を実行するサーバとして構成されており、このサーバに対するクライアントが、他のモジュール、即ち聴覚モジュール20, 顔モジュール30, ステレオモジュール37, モータ制御モジュール40であり、これらのサーバとクライアントとは、互いに非同期で動作する。なお、上記サーバと各クライアントとは、各々、パーソナルコンピュータにより構成されており、更にこれらの各パーソナルコンピュータは、例えばTCP/IPプロトコルの通信環境の下で、相互にLAN(Local Area Network)として構成されている。この場合、好ましくは、データ量の大きいイベントやストリームの通信のためには、ギガビット(Giga bit)のデ

ータ交換が可能な高速ネットワークをロボット視聴覚システム17に適用するのが好ましく、また時刻の同期等の制御用通信のためには中速ネットワークをロボット視聴覚システム17に適用するのが好ましい。このように大きなデータが高速に各パーソナルコンピュータ間を伝送することで、ロボット全体のリアルタイム性及びスケーラビリティを向上させることができる。

【0053】

また、各モジュール20，30，37，40，50は、それぞれ階層的に分散して構成されており、具体的には下位から順次にデバイス層、プロセス層、特徴層、イベント層から構成されている。

【0054】

上記聴覚モジュール20は、デバイス層としてのマイク16と、プロセス層としてのピーク抽出部21、音源定位部22、音源分離部23及びアクティブ方向通過型フィルタ23aと、特徴層（データ）としてのピッチ24、音源水平方向25と、イベント層としての聴覚イベント生成部26と、さらにプロセス層としての音声認識部27及び会話部28と、から構成されている。

【0055】

ここで、聴覚モジュール20は、図5に示すように作用する。即ち、図5において、聴覚モジュール20は、例えば48kHz、16ビットでサンプリングされたマイク16からの音響信号を、符号X1で示すようにFFT（高速フーリエ変換）により周波数解析して、符号X2で示すように左右のチャンネル毎にスペクトルを生成する。そして、聴覚モジュール20は、ピーク抽出部21により左右のチャンネル毎に一連のピークを抽出して、左右のチャンネルで同じか類似のピークをペアとする。

【0056】

ここで、ピーク抽出は、 (α) パワーがしきい値以上で且つ (β) ローカルピークであって、 (γ) 低周波ノイズとパワーの小さい高周波帯域をカットするため例えば90Hz乃至3kHzの間の周波数であるという、3つの条件 $(\alpha \sim \gamma)$ を満たすデータのみを透過させる帯域フィルタを使用して行なわれる。このしきい値は、周囲の暗騒音を計測して、さらに感度パラメータ、例えば10dBを

加えた値として定義される。

【0057】

そして、聴覚モジュール20は、各ピークが調波構造を有していることを利用して音源分離を行う。具体的には、音源分離部23は、周波数の低い方から順に調波構造を有するローカルピークを抽出して、この抽出されたピークの集合を一つの音とみなす。このようにして、音源毎の音響信号が混合音からそれぞれ分離される。音源分離の際、聴覚モジュール20の音源定位部22は、符号X3で示すように、音源分離部23にて分離された各音源毎の音響信号に対して、左右のチャンネルから同じ周波数の音響信号を選択して、IPD（相互位相差）及びIID（相互強度差）を計算する。なお、この計算は、例えば5度毎に行われる。そして、音源定位部22は、計算結果をアクティブ方向通過型フィルタ23aに出力する。

【0058】

これに対して、アクティブ方向通過型フィルタ23aは、アソシエーションモジュール50にて算出されたアソシエーションストリーム59の方向 θ に基づいて、符号X4で示すように、IPDの理論値IPD($=\Delta\phi'(\theta)$)を生成すると共に、IIDの理論値IID($=\Delta\rho'(\theta)$)を計算する。なお、方向 θ は、顔定位（顔イベント39）とステレオ視覚（ステレオ視覚イベント39a）と音源定位（聴覚イベント29）とに基づいて、アソシエーションモジュール50におけるリアルタイムトラッキング（符号X3'）による算出結果である。

【0059】

ここで、理論値IPDと理論値IIDの各計算は、以下に説明する聴覚エピポーラ幾何を利用して行われ、具体的にはロボット10の正面を0度と設定し、±90度の範囲で理論値IPD及び理論値IIDが計算される。ここで、上記聴覚エピポーラ幾何は、HRTFを使用せずに音源の方向情報を得るために必要である。ステレオ視覚研究においては、エピポーラ幾何が最も一般的な定位法の一つであり、聴覚エピポーラ幾何は視覚におけるエピポーラ幾何の聴覚への応用である。そして、聴覚エピポーラ幾何が幾何学的関係を利用して方向情報を得るので、HRTFを不要にすることができる。

【0060】

上記聴覚エピポーラ幾何においては、音源が無限遠にあると仮定し、 $\Delta\phi$ 、 θ 、 f 、 v をそれぞれ I P D、音源方向、周波数、音速とし、 r をロボット頭部を球形とみなした場合の半径とすると、以下の式（1）

【数1】

$$\Delta\phi = \frac{2\pi f}{v} \times r(\theta + \sin\theta)$$

により表わされる。

【0061】

他方、FFT（高速フーリエ変換）により得られた一対のスペクトルに基づいて、各サブバンドの I P D $\Delta\phi'$ 及び I I D $\Delta\rho'$ を、以下の式（2）、（3）により計算する。

【数2】

$$\Delta\phi' = \arctan\left(\frac{\Im[S_{p_l}]}{\Re[S_{p_l}]}\right) - \arctan\left(\frac{\Im[S_{p_r}]}{\Re[S_{p_r}]}\right)$$

【数3】

$$\Delta\rho' = 20 \log_{10} \left(\frac{|S_{p_l}|}{|S_{p_r}|} \right)$$

ここで、 S_{p_l} 、 S_{p_r} は、それぞれある時刻に左右のマイク 16a、16b から得られたスペクトルである。

【0062】

さらに、アクティブ方向通過型フィルタ 23a は、符号 X 7 で示す通過帯域関数に従って、前記ストリーム方向 θ_S から、 θ_S に対応するアクティブ方向通過型フィルタ 23a の通過帯域 $\delta(\theta_S)$ を選択する。ここで、通過帯域関数は、図 5 の X 7 に示すように、ロボットの正面方向 ($\theta = 0$ 度) で感度が最大となり、側方で感度が低下することから、 $\theta = 0$ 度で最小値をとり、側方でより大きくなるような関数である。これは、正面方向で定位の感度が最大になり、左右に角度が大きくなるにつれて感度が低下するという聴覚特性を再現するためのものである。なお、正面方向で定位の感度が最大になることは、哺乳類の目の構造に見

られる中心窓にならって聴覚中心窓と呼ぶ。この聴覚中心窓に関して、人間の場合には、正面の定位の感度が±2度程度であり、左右90度付近にて±8度程度とされている。

【0063】

そして、アクティブ方向通過型フィルタ23a'は、選択した通過帯域 δ (θ_S) を使用して、 θ_L から θ_H の範囲にある音響信号を抽出する。尚、 $\theta_L = \theta_S - \delta$ (θ_S) , $\theta_H = \theta_S + \delta$ (θ_S) と定義する。

【0064】

また、アクティブ方向通過型フィルタ23aは、符号X5で示すように、ストリーム方向 θ_S を頭部伝達関数(HRTF)を利用して、 θ_L 及び θ_H におけるIPD及びIIDの理論値IPD (= $\Delta\phi_H(\theta_S)$) と IID (= $\Delta\rho_H(\theta_S)$) とを、即ち抽出すべき音源の方向を推定する。そして、アクティブ方向通過型フィルタ23aは、音源方向 θ に対して聴覚エピポーラ幾何に基づいて各サブバンド毎に計算されたIPD (= $\Delta\phi_E(\theta)$) 及び IID (= $\Delta\rho_E(\theta)$) と、HRTFに基づいて得られたIPD (= $\Delta\phi_H(\theta)$) 及び IID (= $\Delta\rho_H(\theta)$) とに基づいて、符号X6で示すように、前述した通過帯域 δ (θ) により決定される角度 θ_L から θ_H の角度範囲で、抽出されたIPD (= $\Delta\phi_E$) 及び IID (= $\Delta\rho_E$) が以下の条件を満たすようなサブバンドを集めることとする。

【0065】

ここで、周波数 f_{th} は、フィルタリングの判断基準としてIPDまたはIIDを採用する閾値であって、IPDによる定位が有効である周波数の上限を示す。なお、周波数 f_{th} は、ロボット10のマイク間距離に依存し、本実施形態においては例えば1500Hz程度である。

【0066】

即ち、

【数4】

$$\begin{aligned} f < f_{th} & : \quad \Delta\phi_E(\theta_l) \leq \Delta\phi' \leq \Delta\phi_E(\theta_h) \\ f \geq f_{th} & : \quad \Delta\rho_H(\theta_l) \leq \Delta\rho' \leq \Delta\rho_H(\theta_h) \end{aligned}$$

【0067】

これは、所定周波数 f_{th} 未満の周波数で、 H R T F による I P D の通過帯域 $\delta(\theta)$ の範囲内に I P D ($= \Delta\phi'$) が在る場合、そして所定周波数 f_{th} 以上の周波数で H R T F による I I D の通過帯域 $\delta(\theta)$ の範囲内に I I D ($= \Delta\rho'$) が在る場合に、サブバンドを集めることを意味している。ここで、一般に低周波数帯域では I P D が大きく影響し、高周波数帯域では I I D が大きく影響し、その閾値である周波数 f_{th} はマイク間距離に依存する。

【0068】

そして、アクティブ方向通過型フィルタ 23a は、このようにして集めたサブバンドから音響信号を再合成して、波形を構築することにより、符号 X 8 で示すように、パスーサブバンド方向を生成し、符号 X 9 で示すように、各サブバンド毎にフィルタリングを行なって、符号 X 10 で示す逆周波数変換 IFFT (逆フーリエ変換) により、符号 X 11 で示すように、該当範囲にある各音源からの分離音 (音響信号) を抽出する。

【0069】

上記音声認識部 27 は、図 5 に示すように、自声抑制部 27a と自動認識部 27b とから構成されている。自声抑制部 27a は、聴覚モジュール 20 にて音源定位・音源分離された各音響信号から、後述する対話部 28 のスピーカ 28c から発せられた音声を除去して外部からの音響信号のみを取り出すものである。自動認識部 27b は、図 6 に示すように、音声認識エンジン 27c と音響モデル 27d とセレクタ 27e とから構成されており、この音声認識エンジン 27c としては、例えば京都大学で開発された「Juri an」という音声認識エンジンを利用でき、これにより各話者が発話した単語を認識することができるようになっている。

【0070】

図 6において、自動認識部 27b は、例として男性 2 人 (話者 A, C) と女性 1 人 (話者 B) の三人の話者の認識を行なうように構成されている。このために自動認識部 27b には、各話者の各方向毎にそれぞれ音響モデル 27d が備えられている。図 6 の場合には、音響モデル 27d は、3 人の各話者 A, B, C に関してそれぞれ各話者が発した音声とその方向とを組み合わせて成り、複数種類、

この場合9種類の音響モデル27dが備えられている。

【0071】

音声認識エンジン27cは、並列に9つの音声認識プロセスを実行し、その際に上記9つの音響モデル27dが用いられる。具体的には、音声認識エンジン27cは、それぞれ互いに並列的に入力された音響信号に対して、上記9つの音響モデル27dを用いて音声認識プロセスを実行する。そして、これらの音声認識結果がセレクタ27eに出力される。

【0072】

上記セレクタ27eは、各音響モデル27dからのすべての音声認識プロセス結果を統合して、例えば多数決により最も信頼性が高い音声認識プロセス結果を判断して、その音声認識結果を出力する。

【0073】

ここで、特定話者の音響モデル27dに対する単語認識率を具体的な実験により説明する。

先ず、3m×3mの部屋内において、3つのスピーカをロボット10から1mの位置に且つロボットから0度及び±60度の方向に置く。次に、音響モデル用の音声データとして、男性2名、女性1名が各々発話した、色、数字、食べ物のような150語の単語の音声をスピーカから出力して、ロボット10のマイク16a、16bで集音する。なお、各単語の集音に当たり、一つのスピーカのみからの音声、二つのスピーカから同時に输出される音声、そして三つのスピーカから同時出力される音声、として、各単語に対して3つのパターンを録音する。そして、録音した音声信号に対して前述したアクティブ方向通過型フィルタ23aによって音声分離して各音声データを抽出し、話者及び方向毎に整理して、音響モデルのトレーニングセットを作成する。

【0074】

そして、各音響モデル27dには、トライフォンを使用して、各トレーニングセット毎に、HTK (Hidden Markov Model) ツールキット27fを使用して、各話者の各方向毎に計9種類の音声認識用の音声データを作成した。

【0075】

このようにして得られた音響モデル用音声データを使用して、特定話者の音響モデル27dに対する単語認識率を実験により調べたところ、図7に示す結果が得られた。図7は、横軸に方向を、縦軸に単語認識率を示すグラフであり、符号Pは本人（話者A）の音声、符号Qは他者（話者B, C）の音声の場合を示す。話者Aの音響モデルでは、話者Aがロボット10の正面に位置している場合（図7（A））には、正面（0度）にて80%以上の単語認識率となり、また話者Aが右方60度または左方-60度に位置する場合、それぞれ図7（B）又は図7（C）に示すように、話者よりも方向の違いによる認識率の低下が少なく、特に話者も方向もっている場合には、80%以上の単語認識率となることが分かった。

【0076】

この結果を考慮して、音声認識の際に、音源方向が既知であることを利用して、セレクタ27eは、以下の式（5）により与えられるコスト関数V（pe）を統合のために使用する。

【数5】

$$V(p_e) = \left(\sum_d r(p_e, d) \cdot v(p_e, d) + \sum_d r(p, d_e) \cdot v(p, d_e) - r(p_e, d_e) \right) \cdot P_v(p_e)$$

$$v(p, d) = \begin{cases} 1 & \text{if } \text{Res}(p, d) = \text{Res}(p_e, d_e) \\ 0 & \text{if } \text{Res}(p, d) \neq \text{Res}(p_e, d_e) \end{cases}$$

【0077】

ここで、r（p, d），Res（p, d）をそれぞれ話者pと方向dの音響モデルを使用した場合の単語認識率と入力音声に対する認識結果と定義し、deをリアルタイムトラッキングによる音源方向とし、さらにpeを評価対象の話者とする。

【0078】

上記Pv（pe, de）は顔認識モジュールで生成される確率であり、顔認識ができない場合には常に1.0とする。そして、セレクタ27eは最も大きいコスト関数V（pe）を有する話者peと認識結果Res（p, d）を出力する。

その際、セレクタ27eは、顔モジュール30からの顔認識による顔イベント39を参照することにより、話者を特定することができるので、音声認識のロバスト性を向上させることができる。

【0079】

なお、コスト関数V(p_e)の最大値が1.0以下または二番目に大きい値と近い場合には、音声認識が失敗または一つの候補に絞りきれなかったことにより音声認識ができないと判断して、その旨を後述する対話部28に出力する。上記対話部28は、対話制御部28aと音声合成部28bとスピーカ28cとから構成されている。上記対話制御部28aは、後述するアソシエーションモジュール60により制御されることにより、音声認識部27からの音声認識結果、即ち話者p_eと認識結果R_{es}(p, d)に基づいて、対象とする話者に対する音声データを生成し、音声合成部28bに出力する。上記音声合成部28bは、対話制御部28aからの音声データに基づいてスピーカ28cを駆動して、音声データに対応する音声を発する。

【0080】

これにより、対話部28は音声認識部27からの音声認識結果に基づいて、例えば話者Aが好きな数字として「1」と言った場合に、ロボット10が当該話者Aに正対した状態で、当該話者Aに対して「Aさんは「1」と言いました」というように音声を発することになる。

【0081】

なお、対話部28は、音声認識部27から音声認識ができなかった旨が出力された場合には、ロボット10が当該話者Aに正対した状態で、当該話者Aに対して、「あなたは「2ですか？4ですか？」と質問して、再度話者Aの回答について音声認識を行なうようになっている。この場合、話者Aに対してロボット10が正対していることから、音声認識の精度がより一層向上することになる。

【0082】

このようにして、聴覚モジュール20は、マイク16からの音響信号に基づいて、ピッチ抽出、音源の分離及び定位から少なくとも一人の話者を特定（話者同定）してその聴覚イベントを抽出し、ネットワークを介してアソシエーションモ

ジユール50に対して送信すると共に、各話者の音声認識を行なって対話部28により音声認識結果を話者に対して音声により確認するようになっている。

【0083】

ここで、実際には、音源方向 θ_s が時間tの関数であることから、特定音源を抽出し続けるためには時間方向の連続性を考慮する必要があるが、上述したように、リアルタイムトラッキングからのストリーム方向 θ_s により、音源方向を得るようにしている。

【0084】

これによって、リアルタイムトラッキングにて、すべてのイベントをストリームという時間的流れを考慮した表現で表わしているので、同時に複数の音源が存在したり、音源やロボット自身が移動する場合でも、一つのストリームに注目することによって特定音源からの方向情報を連続的に得ることができる。さらに、ストリームは視聴覚のイベントを統合するためにも使用しているので、顔イベントを参照して聴覚イベントにより音源定位を行なうことにより、音源定位の精度が向上することになる。

【0085】

上記顔モジュール30は、デバイス層としてのカメラ15と、プロセス層としての顔発見部31、顔識別部32、顔定位部33と、特徴層（データ）としての顔ID34、顔方向35と、イベント層としての顔イベント生成部36と、から構成されている。

【0086】

これにより、顔モジュール30は、カメラ15からの画像信号に基づいて、顔発見部31により例えば肌色抽出により各話者の顔を検出し、顔識別部32にて前もって登録されている顔データベース38により検索して、一致した顔があった場合、その顔ID34を決定して当該顔を識別すると共に、顔定位部33により当該顔方向35を決定（定位）する。

【0087】

ここで、顔モジュール30は、顔発見部31が画像信号から複数の顔を見つけた場合、各顔について上記処理、即ち識別及び定位そして追跡を行なう。その際

、顔発見部31により検出された顔の大きさ、方向及び明るさがしばしば変化するので、顔発見部31は顔領域検出を行なって、肌色抽出と相関演算に基づくパターンマッチングの組合せによって200m秒以内に複数の顔を正確に検出できるようになっている。

【0088】

顔定位部33は、二次元の画像平面における顔位置を三次元空間に変換し、三次元空間における顔位置を、方位角 θ 、高さ ϕ 及び距離 r のセットとして得る。そして、顔モジュール30は、各顔毎に、顔ID（名前）34及び顔方向35から、顔イベント生成部36により顔イベント39を生成して、ネットワークを介してアソシエーションモジュール50に対して送信するようになっている。

【0089】

上記顔ステレオモジュール37は、デバイス層としてのカメラ15と、プロセス層としての視差画像生成部37a、目標抽出部37bと、特徴層（データ）としての目標方向37cと、イベント層としてのステレオイベント生成部37dとから構成されている。これにより、ステレオモジュール37は、カメラ15からの画像信号に基づいて視差画像生成部37aによって双方のカメラ15の画像信号から視差画像を生成する。次いで、目標抽出部37bが、視差画像を領域分割し、その結果、縦に長い物体が発見されれば、目標抽出部37bはそれを人物候補として抽出し、その目標方向37cを決定（定位）する。ステレオイベント生成部37dは、目標方向37cに基づいてステレオイベント39aを生成し、ネットワークを介してアソシエーションモジュール50に対して送信するようになっている。

【0090】

上記モータ制御モジュール40は、デバイス層としてのモータ41及びポテンショメータ42と、プロセス層としてのPWM制御回路43、AD変換回路44及びモータ制御部45と、データである特徴層としてのロボット方向46と、イベント層としてのモータイベント生成部47とから構成されている。これにより、モータ制御モジュール40においては、モータ制御部45がアテンション制御モジュール57（後述）からの指令に基づいてPWM制御回路43を介してモー

タ41を駆動制御する。また、モータ41の回転位置をポテンショメータ42により検出する。この検出結果は、AD変換回路44を介してモータ制御部45に送られる。そして、モータ制御部45は、AD変換回路44から受け取った信号からロボット方向46を抽出する。モータイベント生成部47は、ロボット方向46に基づいて、モータ方向情報から成るモータイベント48を生成して、ネットワークを介してアソシエーションモジュール50に対して送信するようになっている。

【0091】

上記アソシエーションモジュール50は、上述した聴覚モジュール20、顔モジュール30、ステレオモジュール37、モータ制御モジュール40に対して、階層的に上位に位置付けられており、各モジュール20、30、37、40のイベント層の上位であるストリーム層を構成している。具体的には、アソシエーションモジュール50は、聴覚モジュール20、顔モジュール30、ステレオモジュール37及びモータ制御モジュール40からの非同期イベント51、即ち聴覚イベント29、顔イベント39、ステレオイベント39a及びモータイベント48を同期させて聴覚ストリーム53、顔ストリーム54、ステレオ視覚ストリーム55を生成する絶対座標変換部52と、各ストリーム53、54、55を関連付けてアソシエーションストリーム59を生成し、あるいはこれらストリーム53、54、55の関連付けを解除する関連付け部56と、さらにアテンション制御モジュール57と、ビューア58とを備えている。

【0092】

上記絶対座標変換部52は、聴覚モジュール20からの聴覚イベント29、顔モジュール30からの顔イベント39、ステレオモジュール37からのステレオイベント39aに、モータ制御モジュール40からのモータイベント48を同期させると共に、聴覚イベント29、顔イベント39及びステレオイベント39aに関して、同期させたモータイベントによって、その座標系を絶対座標系に変換することにより、聴覚ストリーム53、顔ストリーム54及びステレオ視覚ストリーム55を生成する。その際、上記絶対座標変換部52は、同一話者の聴覚ストリーム、顔ストリーム及びステレオ視覚ストリームに接続することによって、

聴覚ストリーム53、顔ストリーム54及びステレオ視覚ストリーム55を生成する。

【0093】

また、関連付け部56は、聴覚ストリーム53、顔ストリーム54、ステレオ視覚ストリーム55に基づいて、これらのストリーム53、54、55の時間的つながりを考慮してストリームを関連付け、あるいは関連付けを解除して、アソシエーションストリーム59を生成すると共に、逆にアソシエーションストリーム59を構成する聴覚ストリーム53、顔ストリーム54及びステレオ視覚ストリーム55の結び付きが弱くなれば、関係付けを解除するようになっている。これにより、目標となる話者が移動している場合であっても、当該話者の移動を予測して、その移動範囲となる角度範囲内であれば、上述したストリーム53、54、55の生成を行なうことによって、当該話者の移動を予測して追跡できることになる。

【0094】

また、アテンション制御モジュール57は、モータ制御モジュール40の駆動モータ制御のプランニングのためのアテンション制御を行なうものであり、その際アソシエーションストリーム59、聴覚ストリーム53、顔ストリーム54そしてステレオ視覚ストリーム55の順に優先的に参照して、アテンション制御を行なう。そして、アテンション制御モジュール57は、聴覚ストリーム53、顔ストリーム54及びステレオ視覚ストリーム55の状態とアソシエーションストリーム59の存否に基づいて、ロボット10の動作プランニングを行ない、駆動モータ41の動作の必要があれば、モータ制御モジュール40に対して動作指令としてのモータイベントをネットワークを介して送信する。ここで、アテンション制御モジュール57におけるアテンション制御は、連続性とトリガに基づいており、連続性により同じ状態を保持しようとし、トリガにより最も興味のある対象を追跡しようとして、アテンションを向けるべきストリームを選択して、トラッキングを行なう。

【0095】

このようにして、アテンション制御モジュール57は、アテンション制御を行

なって、モータ制御モジュール40の駆動モータ41の制御のプランニングを行ない、このプランニングに基づいて、モータコマンド64aを生成し、ネットワーク70を介してモータ制御モジュール40に伝送する。これにより、モータ制御モジュール40では、このモータコマンド64aに基づいて、モータ制御部45がPWM制御を行なって、駆動モータ41を回転駆動させて、ロボット10を所定方向に向けるようになっている。

【0096】

ビューア58は、このようにして生成された各ストリーム53, 54, 55, 57をサーバの画面上に表示するものであり、具体的にはレーダチャート58a及びストリームチャート58bにより表示する。レーダチャート58aは、その瞬間におけるストリームの状態、より詳細にはカメラの視野角と音源方向を示し、ストリームチャート58bは、アソシエーションストリーム（太線図示）と聴覚ストリーム、顔ストリーム及びステレオ視覚ストリーム（細線図示）を示している。

【0097】

本発明実施形態による人型ロボット10は以上のように構成されており、以下のように動作する。

まず、ロボット10の前方1mの距離で、斜め左（ $\theta = +60$ 度）、正面（ $\theta = 0$ 度）そして斜め右（ $\theta = -60$ 度）の方向に、それぞれ話者が並んでおり、ロボット10が対話部28により、三人の話者に質問して、各話者が同時に質問に対する回答を行なう。

【0098】

これにより、ロボット10はマイク16が当該話者の音声を拾って、聴覚モジュール20が音源方向を伴う聴覚イベント29を生成して、ネットワークを介してアソシエーションモジュール50に伝送する。これにより、アソシエーションモジュール50は、この聴覚イベント29に基づいて、聴覚ストリーム53を生成する。

【0099】

また、顔モジュール30は、カメラ15による話者の顔の画像を取り込んで、

顔イベント39を生成して、当該話者の顔を顔データベース38により検索し、顔識別を行なうと共に、その結果である顔ID24及び画像をネットワーク70を介してアソシエーションモジュール50に伝送する。なお、当該話者の顔が顔データベース38に登録されていない場合には、顔モジュール30は、その旨をネットワークを介してアソシエーションモジュール50に伝送する。

【0100】

従って、アソシエーションモジュール50は、これらの聴覚イベント29、顔イベント39、ステレオイベント39aに基づいて、アソシエーションストリーム59を生成する。

【0101】

ここで、聴覚モジュール20は、アクティブ方向通過型フィルタ23aにより、聴覚エピポーラ幾何によるIPDを利用して、各音源（話者X, Y, Z）の定位及び分離を行なって、分離音（音響信号）を取り出す。そして、聴覚モジュール20は、その音声認識部27により音声認識エンジン27cを使用して、各話者X, Y, Zの音声を認識してその結果を対話部28に出力する。これにより、対話部28は、音声認識部27により音声認識された前記回答を、それぞれの話者に対してロボット10が正対した状態で発話する。なお、音声認識部27が正しく音声認識できなかった場合には、ロボット10が当該話者に正対した状態で再度質問を繰り返し、その回答に基づいて再度音声認識を行なう。

【0102】

このようにして、本発明実施形態による人型ロボット10によれば、聴覚モジュール20により音源定位・音源分離された分離音（音響信号）に基づいて、音声認識部27が、各話者及び方向に対応する音響モデルを使用して音声認識を行なうことにより同時に発話する複数の話者の音声を音声認識することができる。

【0103】

以下に、音声認識部27の動作を実験により評価する。

これらの実験においては、図8に示すように、ロボット10の前方1mの距離で、斜め左（ $\theta = +60$ 度）、正面（ $\theta = 0$ 度）そして斜め右（ $\theta = -60$ 度）の方向に、それぞれ話者X, Y, Zが並んでいる。なお、実験では、話者として

人間の代わりにそれぞれスピーカを置くと共に、その前面に話者の写真を配置している。ここで、スピーカは、音響モデルを作成したときと同じスピーカを使用しており、スピーカから発せられた音声を写真的話者の音声とみなしている。

【0104】

そして、以下のシナリオに基づいて音声認識の実験を行なう。

1. ロボット10が三人の話者X, Y, Zに質問する。
2. 三人の話者X, Y, Zが同時に質問に対する回答を行なう。
3. ロボット10が三人の話者X, Y, Zの混合音声に基づいて、音源定位・音源分離を行ない、さらに各分離音について音声認識を行なう。
4. ロボット10が、順次に各話者X, Y, Zに正対した状態で当該話者の回答を答える。
5. ロボット10は、音声認識が正しくできなかつたと判断したとき、当該話者に正対して再度質問を繰り返し、その回答に基づいて再度音声認識を行なう。

【0105】

上記シナリオによる実験結果の第一の例を図9に示す。

1. ロボット10が「好きな数字は何ですか？」と質問する。（図9（a）参照）
2. 各話者X, Y, Zとしてのスピーカから、同時に1から10までの数字のうちから、任意の数字を読み上げた音声を流す。例えば図9（b）に示すように、話者Xは「2」、話者Yは「1」そして話者Zは「3」と言う。
3. ロボット10は、聴覚モジュール20にて、そのマイク16で集音した音響信号に基づいて、アクティブ方向通過型フィルタ23aにより音源定位・音源分離を行なって、分離音を抽出する。そして、各話者X, Y, Zに対応する分離音に基づいて、各話者別に音声認識部27が9つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。
4. その際、音声認識部27のセレクタ27eが、正面が話者Yであると仮定して音声認識の評価を行ない（図9（c））、続いて正面が話者Xであると仮定して音声認識の評価を行ない（図9（d））、最後に正面が話者Zであると仮定して音声認識の評価を行なう（図9（e））。

5. そして、セレクタ27eが、音声認識結果を統合して、図9（f）に示すように、ロボット正面（ $\theta = 0$ 度）に関して、最も適合の良い話者名（Y）と音声認識結果（「1」）を決定し対話部28に出力する。これにより、図9（g）に示すように、ロボット10が話者Yに正対した状態にて、「Yさんは「1」です。」と答える。

6. 続いて、斜め左（ $\theta = +60$ 度）の方向に関して、上記と同様の処理を行って、図9（h）に示すように、ロボット10が話者Xに正対した状態にて、「Xさんは「2」です。」と答える。更に、斜め右（ $\theta = -60$ 度）の方向に対しても同様の処理を行って、図9（i）に示すように、ロボット10が話者Zに正対した状態にて、「Zさんは「3」です。」と答える。

【0106】

この場合、ロボット10は、各話者X, Y, Zの回答をすべて正しく音声認識することができた。従って、同時発話の場合であっても、ロボット10のマイク16を使用したロボット視聴覚システム17における音源定位・音源分離・音声認識の有効性が示された。

【0107】

なお、図9（j）に示すように、ロボット10が各話者に正対せずに、「Yさんは「1」です。Xさんは「2」です。Zさんは「3」です。合計「6」です。」というように、各話者X, Y, Zの答えた数字の合計も答えるようにしてもよい。

【0108】

図10は、上述したシナリオによる実験結果の第二の例を示している。

1. 図9に示した第一の例と同様にして、ロボット10が「好きな数字は何ですか？」と質問し（図10（a）参照）、各話者X, Y, Zとしてのスピーカから、図10（b）に示すように、話者Xは「2」、話者Yは「1」そして話者Zは「3」という音声が流れる。

2. ロボット10は、同様にして、聴覚モジュール20にて、そのマイク16で集音した音響信号に基づいて、アクティブ方向通過型フィルタ23aにより音源定位・音源分離を行なって分離音を抽出し、各話者X, Y, Zに対応する分離

音に基づいて、各話者別に音声認識部27が9つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。その際、音声認識部27のセレクタ27eは、図10(c)に示すように、正面の話者Yについては、正しく音声認識の評価を行なうことができる。

3. これに対して、+60度に位置する話者Xについて、セレクタ27eは、図10(d)に示すように、「2」であるか「4」であるか決定することができない。

4. 従って、ロボット10は、図10(e)に示すように、+60度に位置する話者Xに正対して、「2ですか？ 4ですか？」と質問する。

5. これに対して、図10(f)に示すように、話者Xであるスピーカから「2」という回答が流れる。この場合、話者Xは、ロボット10の正面に位置していることから、聴覚モジュール20が話者Xの回答について正しく音源定位・音源分離し、音声認識部27が正しく音声認識して、話者名Xと音声認識結果「2」を対話部28に出力する。これにより、ロボット10は、図10(g)に示すように、話者Xに対して「Xさんは「2」です。」と答える。

6. 続いて、話者Zについても同様の処理を行なって、その音声認識結果を話者Zに対して答える。即ち、図10(h)に示すように、ロボット10が話者Zに正対した状態にて、「Zさんは「3」です。」と答える。

【0109】

このようにして、ロボット10は、再質問により、各話者X, Y, Zの回答をすべて正しく音声認識することができた。従って、側方での聴覚中心窓の影響による分離精度の低下による音声認識の曖昧さを、ロボット10が側方の話者に対して正対して再質問することにより解消して、音源分離精度を向上させ、音声認識精度を向上させることができることが示された。

【0110】

なお、図10(i)に示すように、ロボット10が各話者の音声認識を正しく行なった後、「Yさんは「1」です。Xさんは「2」です。Zさんは「3」です。合計「6」です。」というように、各話者X, Y, Zの答えた数字の合計も答えるようにしてもよい。

【0111】

図11は、上述したシナリオによる実験結果の第三の例を示している。

1. この場合も図9に示した第一の例と同様にして、ロボット10が「好きな数字は何ですか？」と質問し（図10（a）参照）、各話者X, Y, Zとしてのスピーカから、図10（b）に示すように、話者Xは「8」、話者Yは「7」そして話者Zは「9」という音声が流れる。

2. ロボット10は、同様にして、聴覚モジュール20にて、そのマイク16で集音した音響信号に基づいて、リアルタイムトラッキング（X3'参照）によるストリーム方向 θ 、そして各話者の顔イベントを参照して、アクティブ方向通過型フィルタ23aにより音源定位・音源分離を行なって分離音を抽出し、各話者X, Y, Zに対応する分離音に基づいて、各話者毎に音声認識部27が9つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。

その際、音声認識部27のセレクタ27eは、正面の話者Yについては、顔イベントに基づいて話者Yである確率が高いことから、各音響モデルによる音声認識結果の統合の際に、図10（c）に示すようにこれを考慮する。これにより、より正確な音声認識を行なうことができる。従って、ロボット10は、図11（d）に示すように、話者Xに対して「Xさんは「7」です。」と答える。

3. これに対して、+60度に位置する話者Xについて、ロボット10が向きを変えて正対すると、このときの正面の話者Xについて、顔イベントに基づいて話者Xである確率が高いので、同様にして、セレクタ27eは、図11（e）に示すようにこれを考慮する。従って、ロボット10は、図11（f）に示すように、話者Xに対して「Yさんは「8」です。」と答える。

4. 続いて、セレクタ27eは、図11（g）に示すように、話者Zについても同様の処理を行なって、その音声認識結果を話者Zに対して答える。即ち、図11（h）に示すように、ロボット10が話者Zに正対した状態にて「Zさんは「9」です。」と答える。

【0112】

このようにして、ロボット10は、各話者毎に正対して、その顔イベントを参

照しながら話者の顔認識に基づいて、各話者X, Y, Zの回答をすべて正しく音声認識することができた。これにより、顔認識により話者が誰であるかを特定することができるので、より精度の高い音声認識を行なうことができる事が示された。特に、特定の環境での利用を前提とするような場合、顔認識によってほぼ100%に近い顔認識精度が得られると、顔認識情報を信頼性の高い情報として利用することができることになり、音声認識部27の音声認識エンジン27cで使用される音響モデル27dの数を削減することができるので、より高速で且つ高精度の音声認識が可能になる。

【0113】

図12は、上述したシナリオによる実験結果の第四の例を示している。

1. ロボット10が「好きなフルーツは何ですか？」と質問し（図12（a）参照）、各話者X, Y, Zとしてのスピーカから、例えば図12（b）に示すように、話者Xは「梨」、話者Yは「スイカ」そして話者Zは「メロン」と言う。
2. ロボット10は、聴覚モジュール20にて、そのマイク16で集音した音響信号に基づいて、アクティブ方向通過型フィルタ23aにより音源定位・音源分離を行なって分離音を抽出する。そして、各話者X, Y, Zに対応する分離音に基づいて、各話者毎に音声認識部27が9つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。
3. その際、音声認識部27のセレクタ27eが、正面が話者Yであると仮定して音声認識の評価を行ない（図12（c））、続いて正面が話者Xであると仮定して音声認識の評価を行ない（図12（d））、最後に正面が話者Zであると仮定して音声認識の評価を行なう（図12（e））。
4. そして、セレクタ27eが、音声認識結果を統合して、図12（f）に示すように、ロボット正面（θ=0度）方向に関して最も適合の良い話者名（Y）と音声認識結果（「スイカ」）を決定し対話部28に出力する。これにより、図9（g）に示すように、ロボット10が話者Yに正対した状態にて、「Yさんは「スイカ」です。」と答える。
5. 続いて、各話者X, Zについても同様の処理を行なって、その音声認識結果を各話者X, Zに対して答える。即ち、図12（h）に示すように、ロボット

10が話者Xに正対した状態にて、「Xさんは「梨」です。」と答え、さらに図12(i)に示すように、ロボット10が話者Zに正対した状態にて「Zさんは「メロン」です。」と答える。

【0114】

この場合、ロボット10は、各話者X, Y, Zの回答をすべて正しく音声認識することができた。従って、音声認識エンジン27cに登録された単語は数字に限ることなく、前もって登録された単語であれば、音声認識可能であることが分かる。ここで、実験に使用した音声認識エンジン27cでは、約150語の単語が登録されている。なお、単語の音節数が多くなると、音声認識率はやや低くなる。

【0115】

上述した実施形態においては、ロボット10は、その上半身が4DOF（自由度）を有するように構成されているが、これに限らず、任意の動作を行なうように構成されたロボットに本発明によるロボット視聴覚システムを組み込むことも可能である。

【0116】

また、上述した実施形態においては、本発明によるロボット視聴覚システムを人型ロボット10に組み込んだ場合について説明したが、これに限らず、大型等の各種動物型ロボットや、その他の形式のロボットに組み込むことも可能であることは明らかである。

【0117】

また、上記説明では、図4に示すようにロボット視聴覚システム17がステレオモジュール37を備える構成例を説明したが、本発明の実施形態に係るロボット視聴覚システムは、ステレオモジュール37を備えずに構成することもできる。この場合、アソシエーションモジュール50は、聴覚イベント29, 顔イベント39及びモータイベント48に基づいて、各話者の聴覚ストリーム53及び顔ストリーム54を生成し、さらにこれらの聴覚ストリーム53及び顔ストリーム54を関連付けてアソシエーションストリーム59を生成するように構成され、アテンション制御モジュール50においては、これらのストリームに基づいてア

テンション制御が行われるように構成される。

【0118】

さらに、上記説明においては、アクティブ方向通過型フィルタ23aは、方向毎に通過帯域幅（パスレンジ）を制御しており、処理する音の周波数によらず通過帯域幅を一定としていた。

ここで、通過帯域δを導出するために、100Hz, 200Hz, 500Hz, 1000Hz, 2000Hz, 100Hzの調波構造音（ハーモニクス）の5つの純音と1つのハーモニクスとを用いて、1音源に対する音源抽出率を調べる実験を行った。なお、音源をロボット正面である0度からロボットの左位置或いは右位置である90度の範囲で10度毎に位置を移動させた。図13～図15は音源を0度から90度の範囲の各位置に設置した場合の音源抽出率を示すグラフであり、この実験結果が示すように、周波数に応じて通過帯域幅を制御することにより、特定の周波数の音の抽出率を向上させることができ、分離精度を向上できる。よって、音声認識率も向上する。従って、上記説明したロボット視聴覚システム17においては、アクティブ方向通過型フィルタ23aのパスレンジが、周波数毎に制御可能に構成されるのが望ましい。

【0119】

【発明の効果】

以上述べたように、この発明によれば、従来の音声認識と比較して、複数の音響モデルを使用することによって、実時間・実環境での正確な音声認識を行なうことが可能である。また、各音響モデルによる音声認識結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断することにより、従来の音声認識に比べて、より一層正確な音声認識を行なうことができる。

【図面の簡単な説明】

【図1】

この発明によるロボット聴覚装置の第一の実施形態を組み込んだ人型ロボットの外観を示す正面図である。

【図2】

図1の人型ロボットの側面図である。

【図3】

図1の入型ロボットにおける頭部の構成を示す概略拡大図である。

【図4】

図1の入型ロボットにおけるロボット視聴覚システムの電気的構成例を示すブロック図である。

【図5】

図4に示すロボット視聴覚システムにおける聴覚モジュールの作用を示す図である。

【図6】

図4のロボット視聴覚システムにおける聴覚モジュールの音声認識部で使用される音声認識エンジンの構成例を示す概略斜視図である。

【図7】

図6の音声認識エンジンによる正面及び左右±60度の方向の話者による音声の認識率を示すグラフであり、(A)は正面の話者、(B)は斜め左+60度の話者そして(C)は斜め右-60度の話者の場合を示している。

【図8】

図4に示すロボット視聴覚システムにおける音声認識実験を示す概略斜視図である。

【図9】

図4のロボット視聴覚システムの音声認識実験の第一の例の結果を順次に示す図である。

【図10】

図4のロボット視聴覚システムの音声認識実験の第二の例の結果を順次に示す図である。

【図11】

図4のロボット視聴覚システムの音声認識実験の第三の例の結果を順次に示す図である。

【図12】

図4のロボット視聴覚システムの音声認識実験の第四の例の結果を順次に示す

図である。

【図13】

本発明の実施形態に係るアクティブ方向通過型フィルタの通過帯域幅を制御した場合の抽出率を示す図であり、(a)は0度、(b)は10度、(c)は20度、(d)は30度の方向に音源がある場合である。

【図14】

本発明の実施形態に係るアクティブ方向通過型フィルタの通過帯域幅を制御した場合の抽出率を示す図であり、(a)は40度、(b)は50度、(c)は60度の方向に音源がある場合である。

【図15】

本発明の実施形態に係るアクティブ方向通過型フィルタの通過帯域幅を制御した場合の抽出率を示す図であり、(a)は70度、(b)は80度、(c)は90度の方向に音源がある場合である。

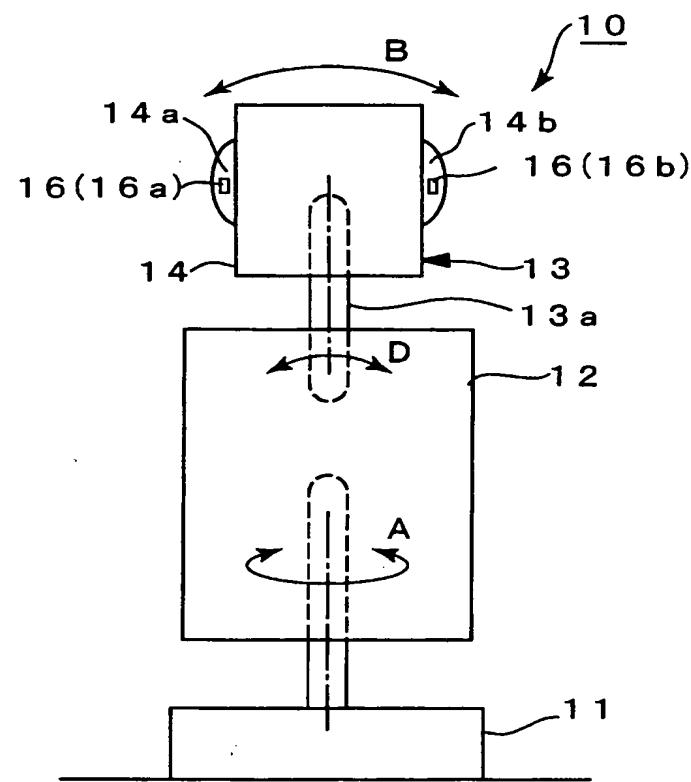
【符号の説明】

- 1 0 人型ロボット
- 1 1 ベース
- 1 2 胴体部
- 1 3 頭部
- 1 4 外装
- 1 5 カメラ（ロボット視覚）
- 1 6, 1 6 a, 1 6 b マイク（ロボット聴覚）
- 1 7 ロボット視聴覚システム
- 2 0 聴覚モジュール
- 2 1 ピーク抽出部
- 2 2 音源定位部
- 2 3 音源分離部
- 2 3 a アクティブ方向通過型フィルタ
- 2 6 聴覚イベント生成部
- 2 7 音声認識部

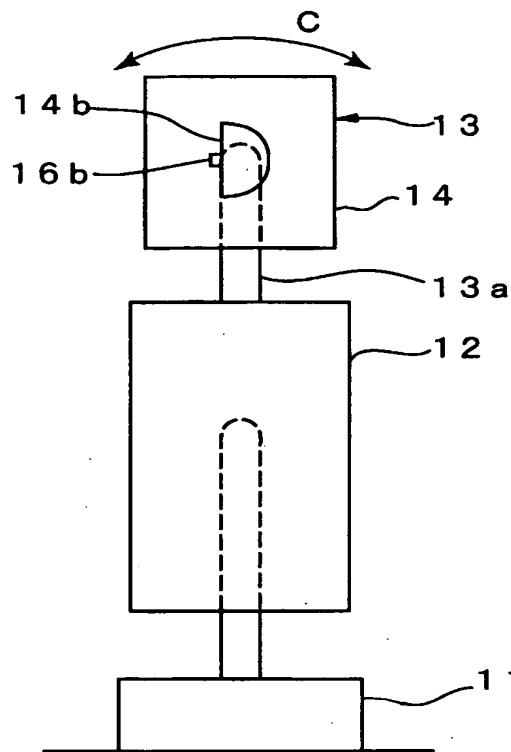
- 27a 自声抑制部
- 27b 自動認識部
- 27c 音声認識エンジン
- 27d 音響モデル
- 27e セレクタ
- 28 対話部
- 30 顔モジュール
- 37 ステレオモジュール
- 40 モータ制御モジュール
- 50 アソシエーションモジュール
- 57 アテンション制御モジュール

【書類名】 図面

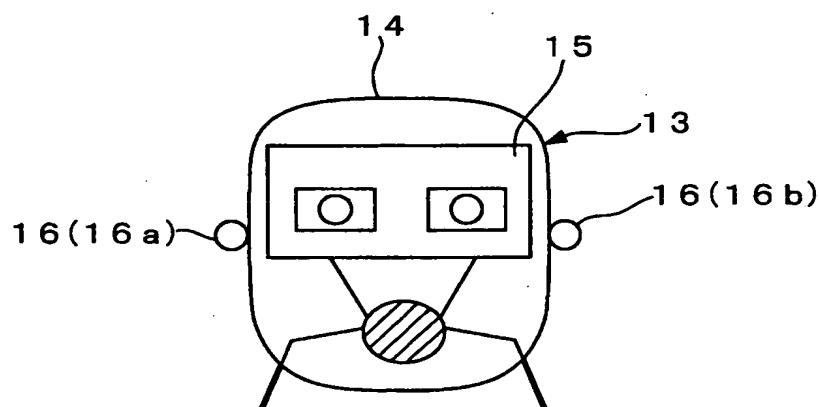
【図1】



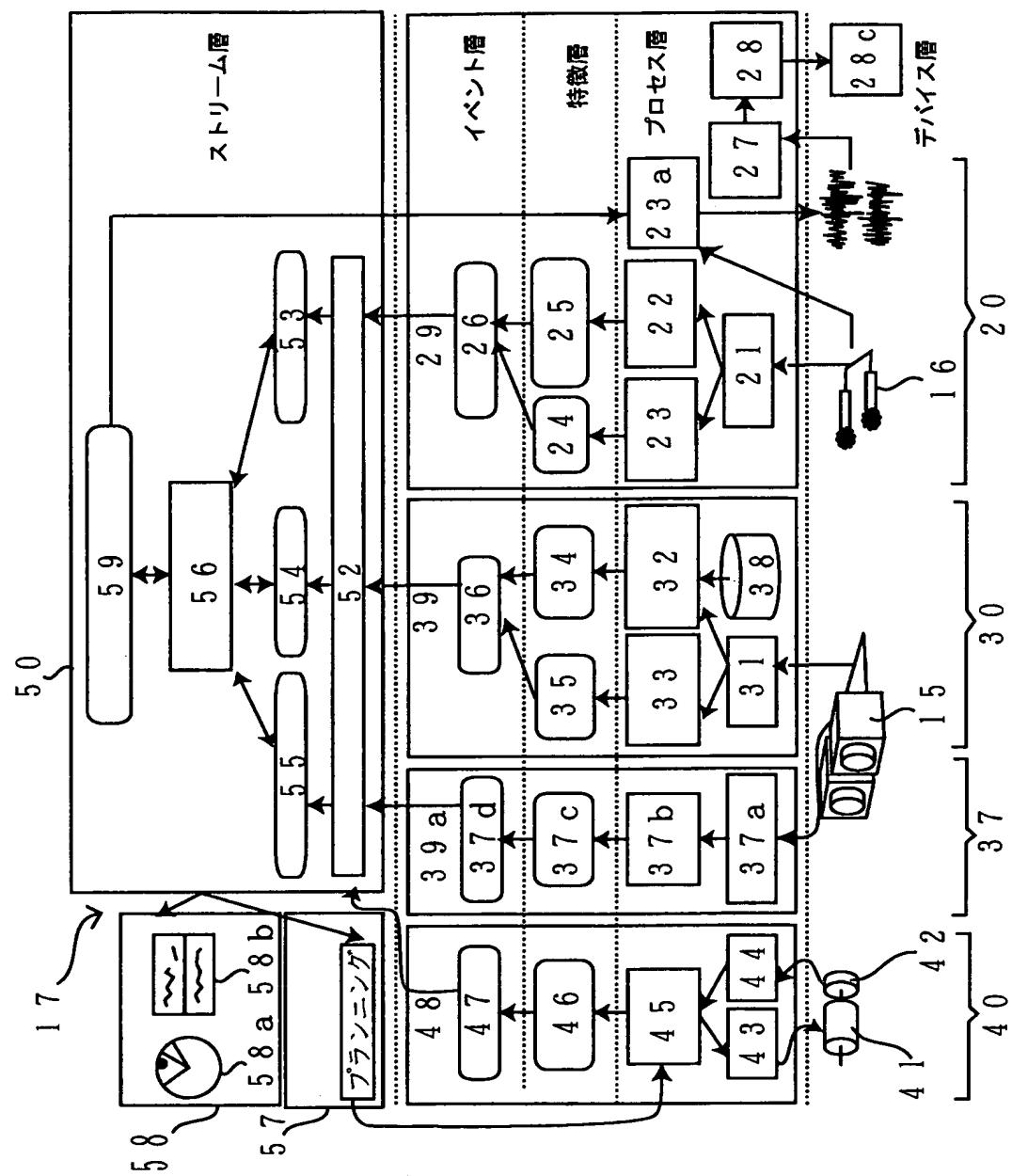
【図2】



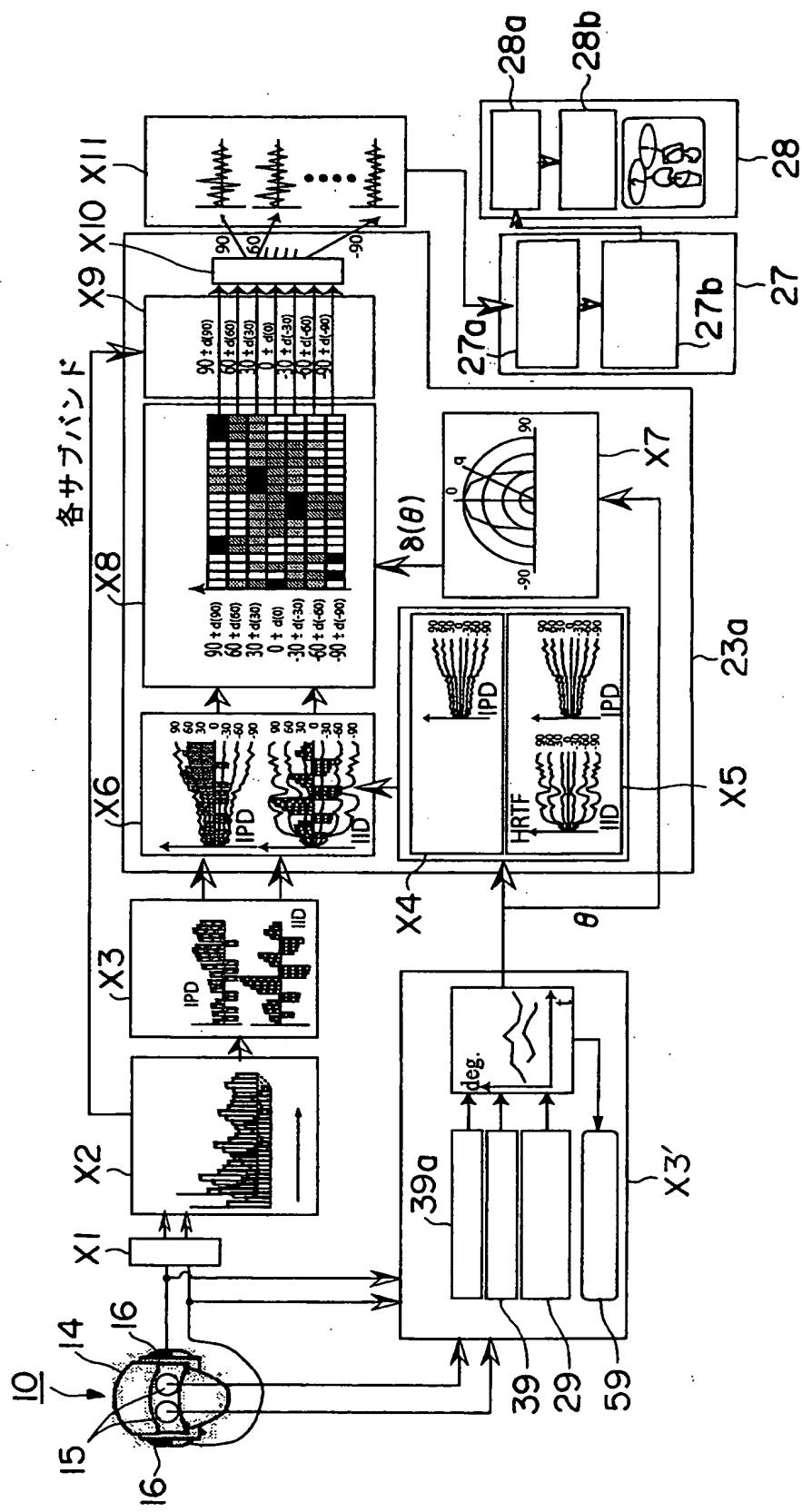
【図3】



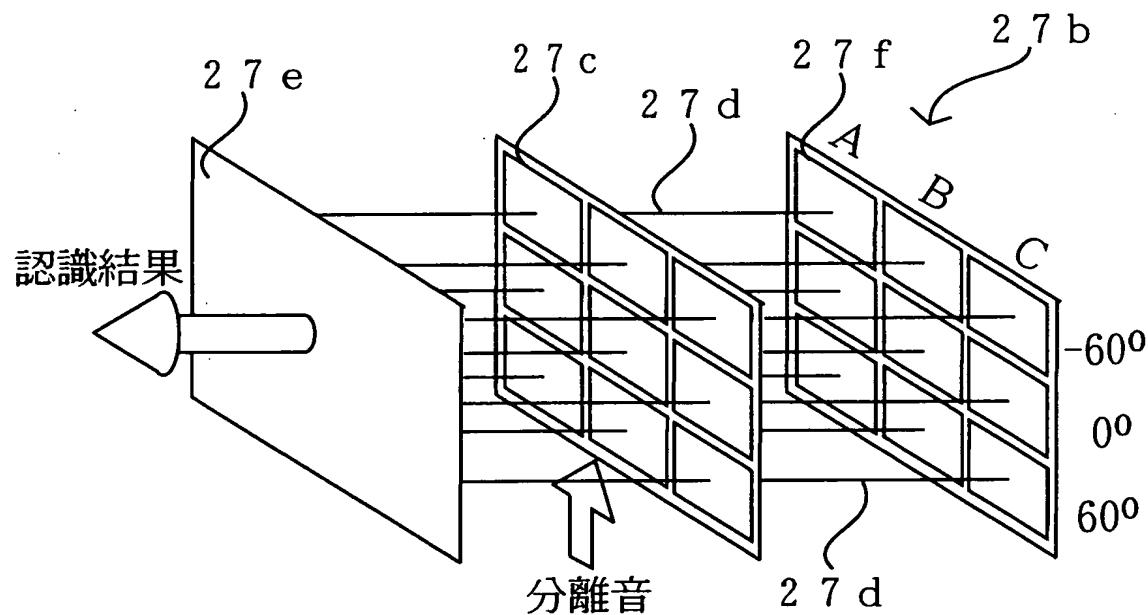
【図4】



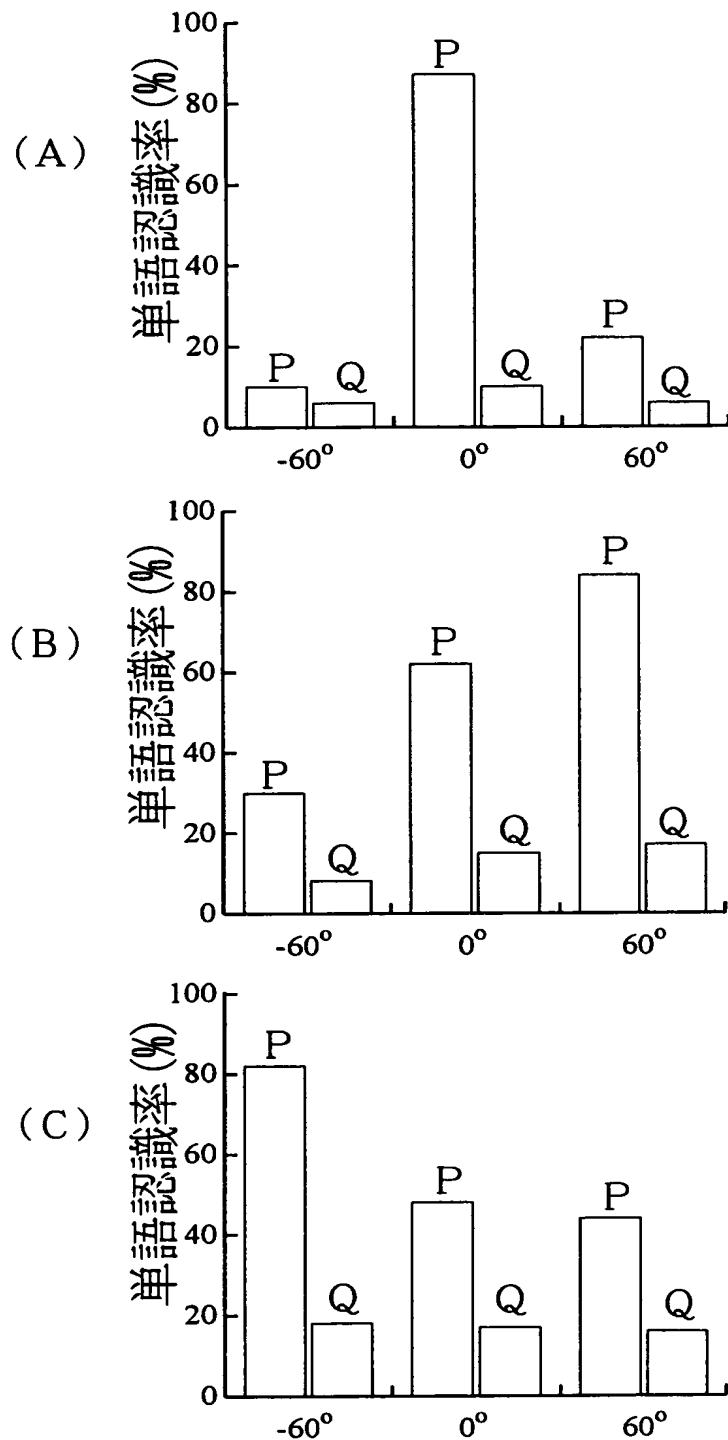
【図5】



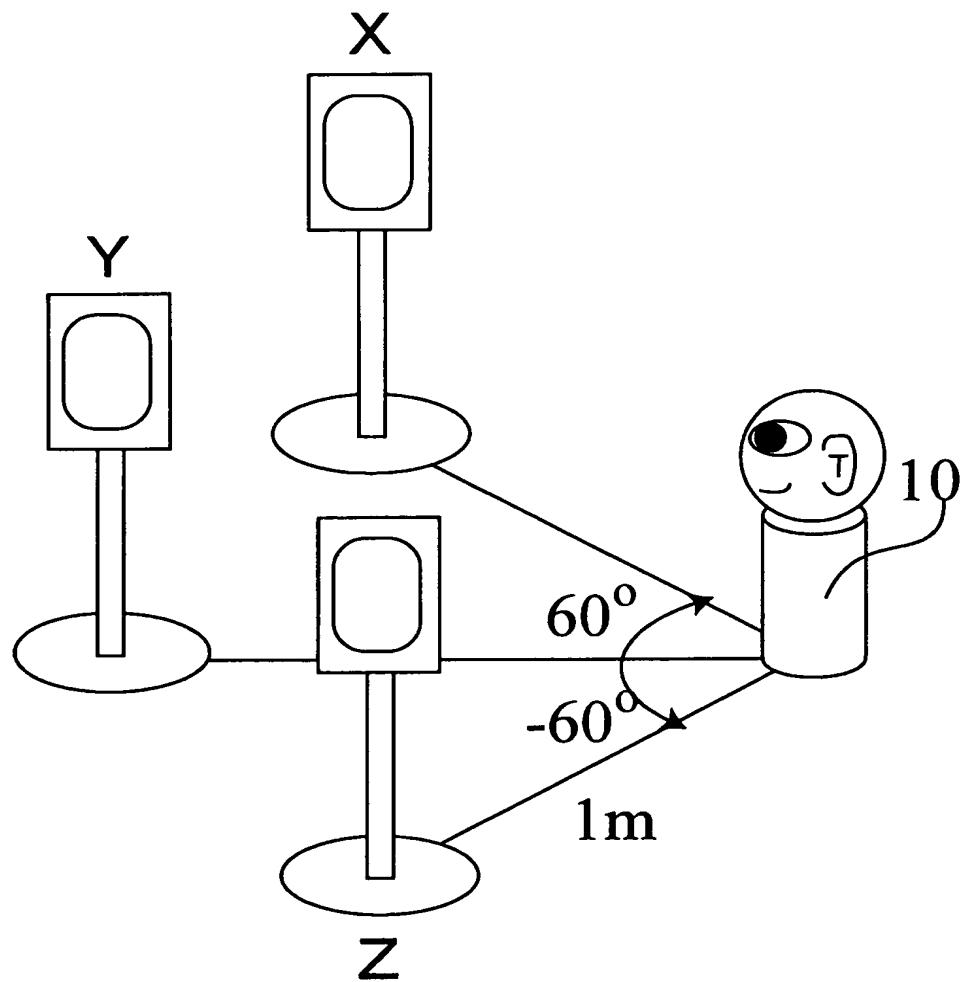
【図6】



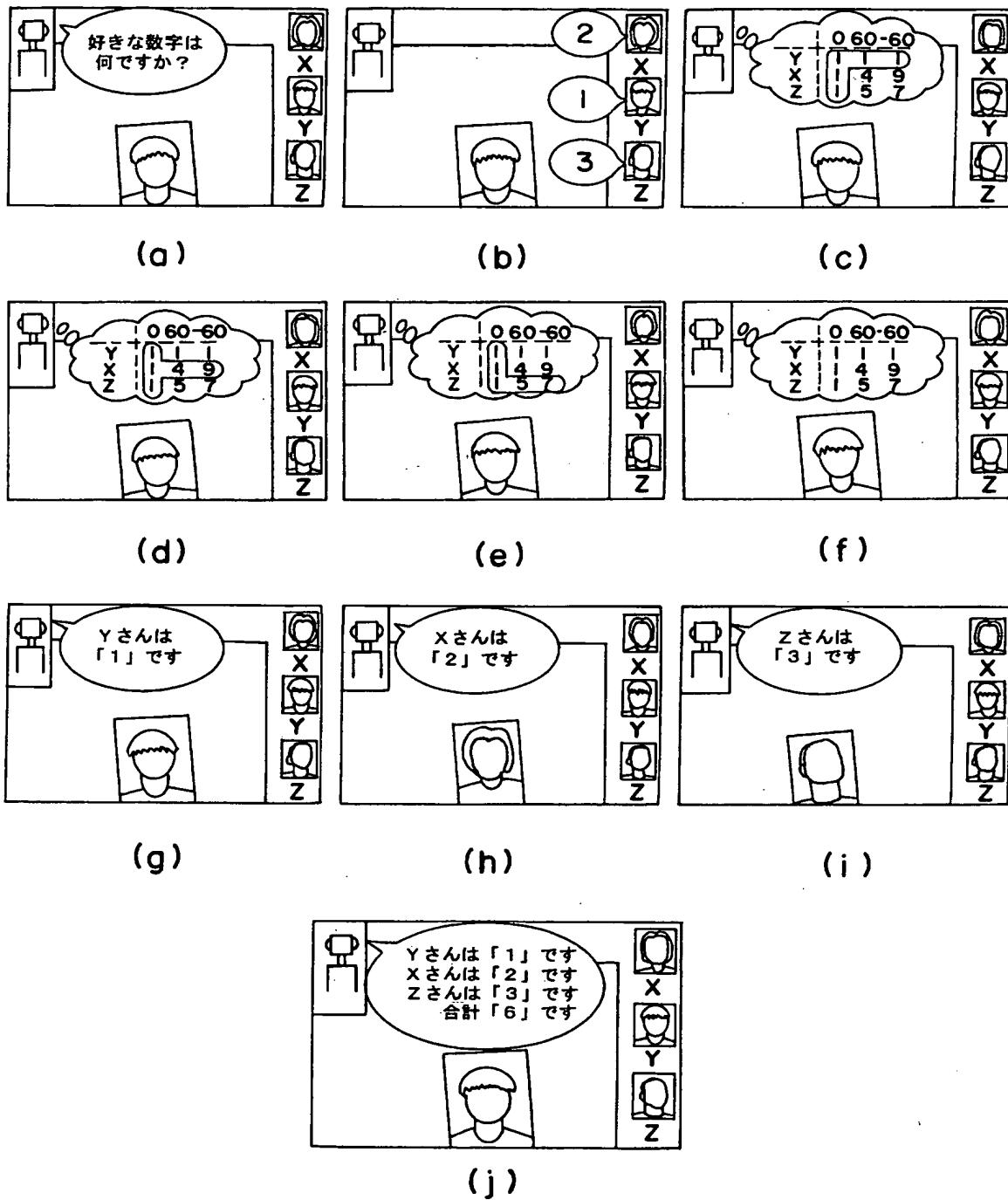
【図 7】



【図8】



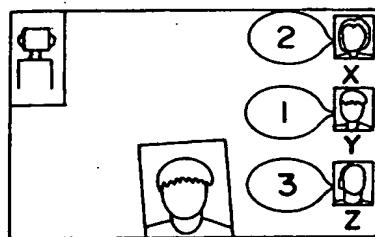
【図9】



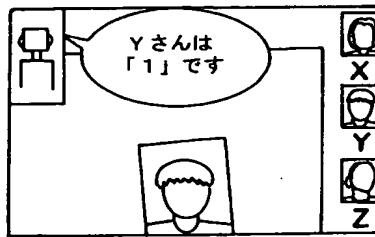
【図10】



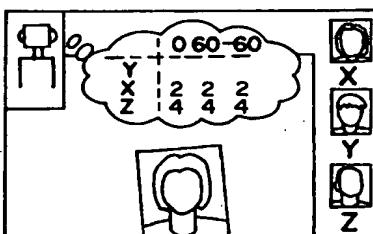
(a)



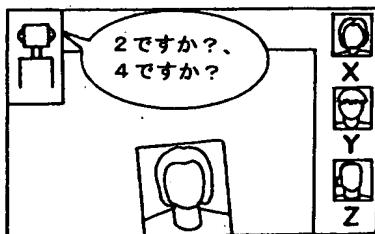
(b)



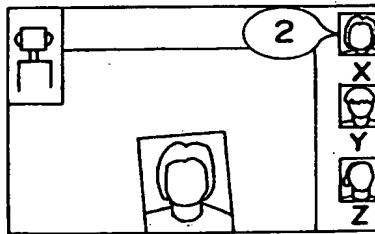
(c)



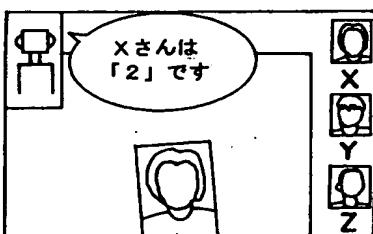
(d)



(e)



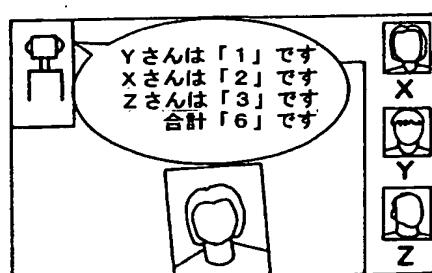
(f)



(g)



(h)

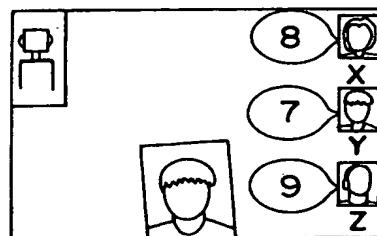


(i)

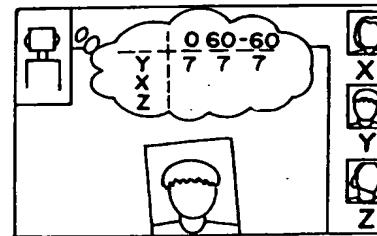
【図11】



(a)



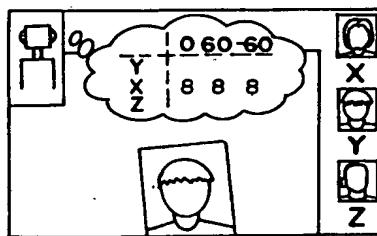
(b)



(c)



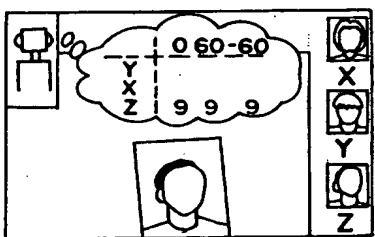
(d)



(e)



(f)



(g)

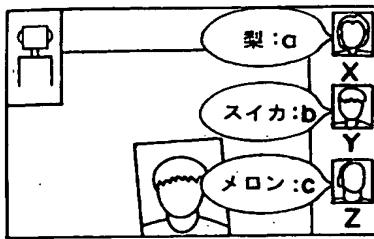


(h)

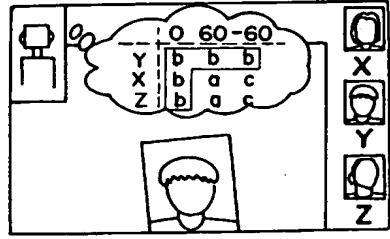
【図12】



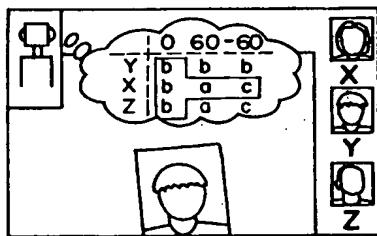
(a)



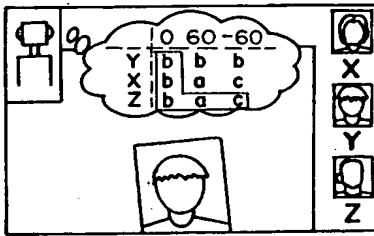
(b)



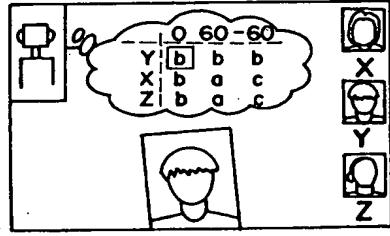
(c)



(d)



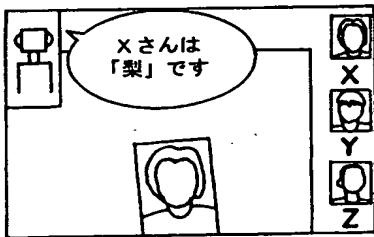
(e)



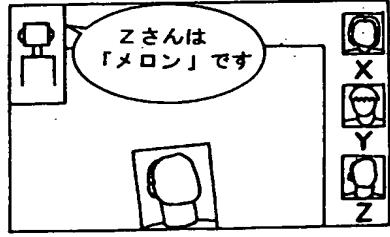
(f)



(g)

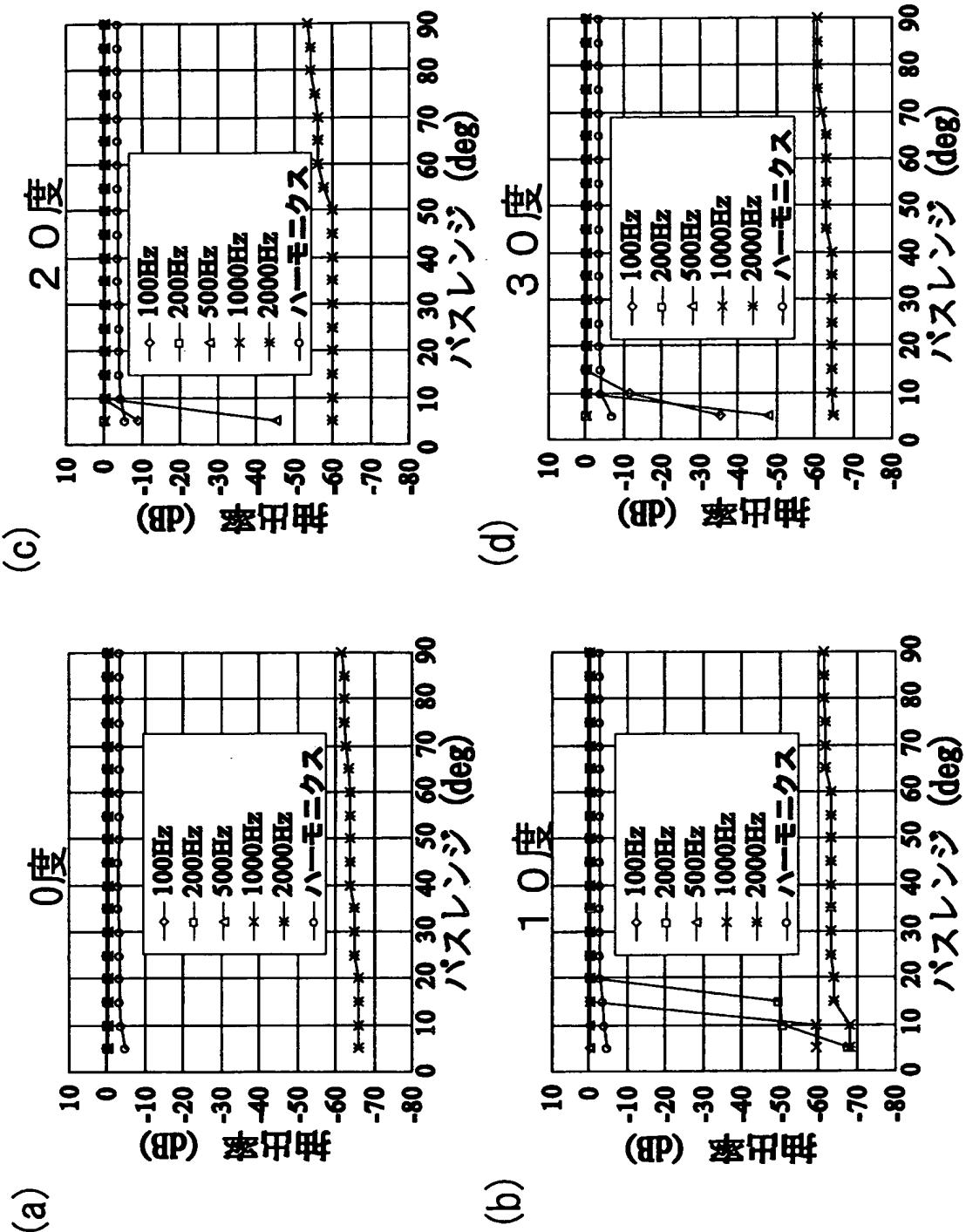


(h)



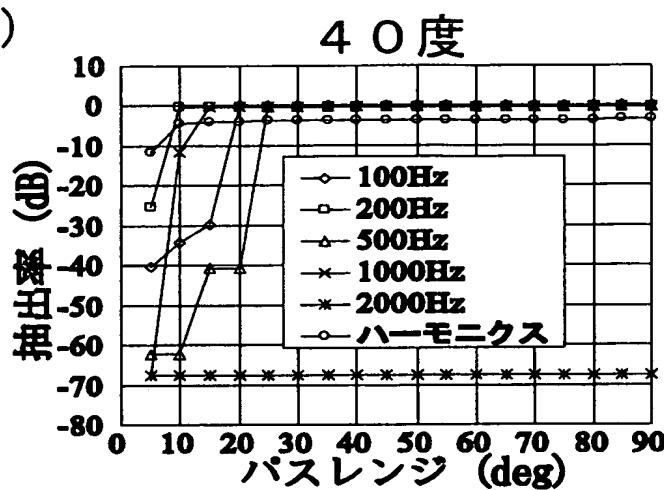
(i)

【図13】

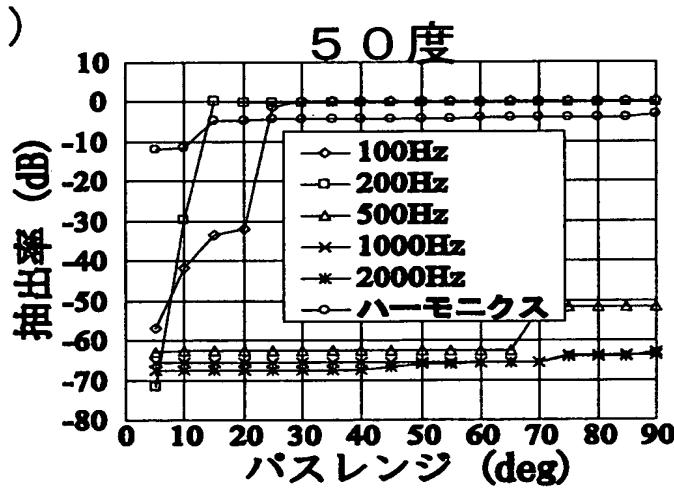


【図14】

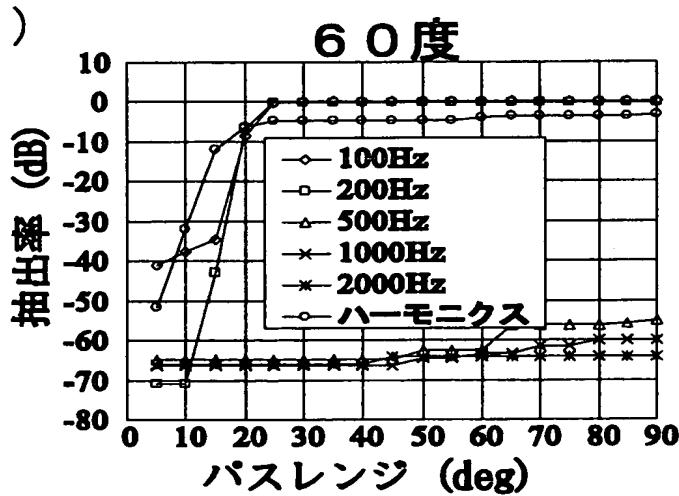
(a)



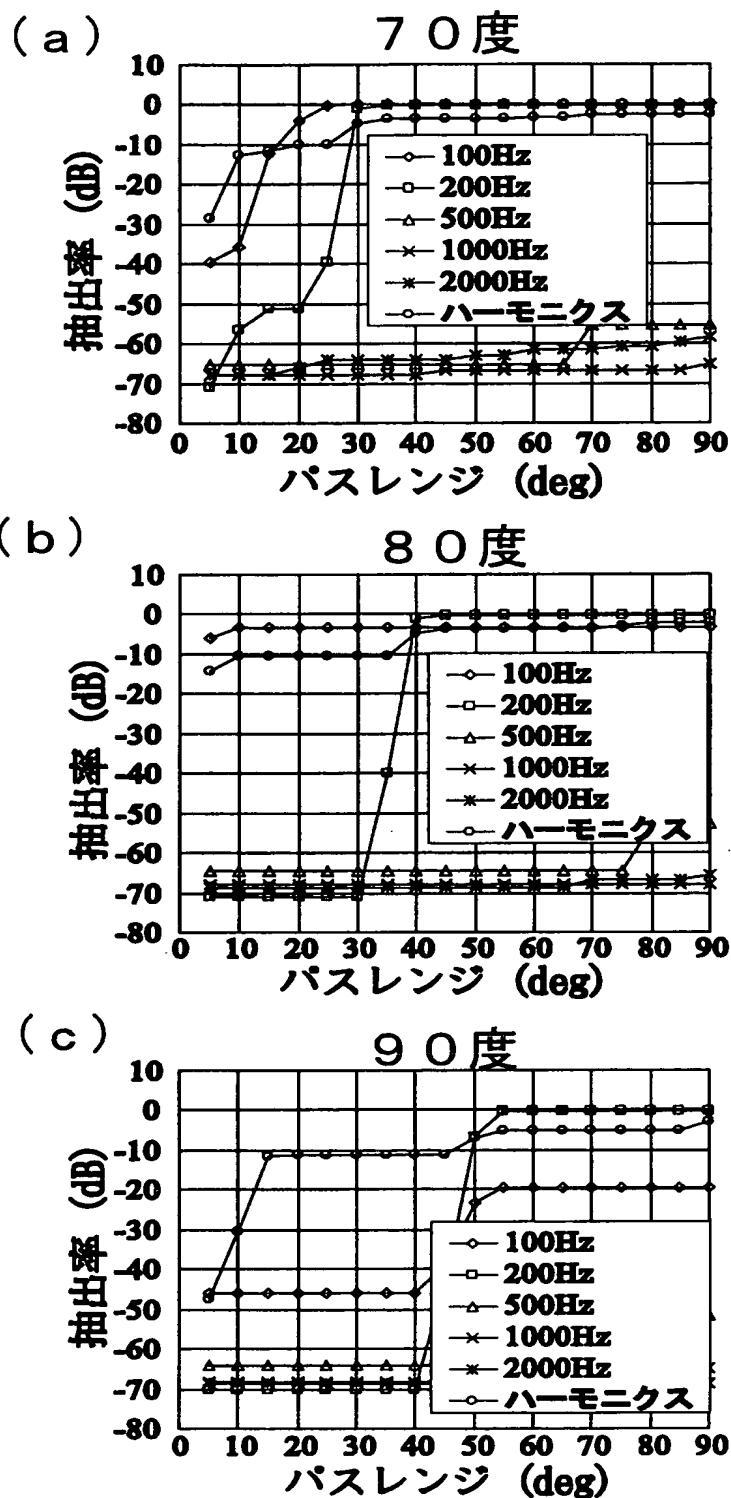
(b)



(c)



【図15】



【書類名】 要約書

【要約】

【課題】 各音源からの分離された音についての認識を行なうようにしたロボット視聴覚システムを提供する。

【解決手段】 聴覚モジュール20，顔モジュール30，ステレオモジュール37，モータ制御モジュール40と、これらの各モジュールを制御するアソシエーションモジュール50とを備え、聴覚モジュールが、複数の音響モデルにより音声認識を行ない、各音響モデルによる音声認識結果をセレクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されている。

【選択図】 図4

【書類名】 出願人名義変更届（一般承継）
【提出日】 平成15年10月31日
【あて先】 特許庁長官 殿
【事件の表示】
 【出願番号】 特願2002-365764
【承継人】
 【識別番号】 503360115
 【住所又は居所】 埼玉県川口市本町四丁目1番8号
 【氏名又は名称】 独立行政法人科学技術振興機構
 【代表者】 沖村 憲樹
 【連絡先】 〒102-8666 東京都千代田区四番町5-3 独立行政法人科学技術振興機構 知的財産戦略室 佐々木吉正 TEL 03-5214-8486 FAX 03-5214-8417

【提出物件の目録】
【物件名】 権利の承継を証明する書面 1
【援用の表示】 平成15年10月31日付提出の特第許3469156号にかかる一般承継による移転登録申請書に添付のものを援用する。
【物件名】 登記簿謄本 1
【援用の表示】 平成15年10月31日付提出の特第許3469156号にかかる一般承継による移転登録申請書に添付のものを援用する。

特願 2002-365764

出願人履歴情報

識別番号 [396020800]

1. 変更年月日 1998年 2月24日

[変更理由] 名称変更

住所 埼玉県川口市本町4丁目1番8号

氏名 科学技術振興事業団

特願 2002-365764

出願人履歴情報

識別番号 [503360115]

1. 変更年月日 2003年10月 1日
[変更理由] 新規登録
住 所 埼玉県川口市本町4丁目1番8号
氏 名 独立行政法人 科学技術振興機構

2. 変更年月日 2004年 4月 1日
[変更理由] 名称変更
住 所 埼玉県川口市本町4丁目1番8号
氏 名 独立行政法人科学技術振興機構

3. 変更年月日 2009年 8月 20日
[変更理由] 住所変更
住 所 東京都千代田区四番町5-3 サイエンスプラザ5F
氏 名 独立行政法人科学技術振興機構

4. 変更年月日 2009年 10月 8日
[変更理由] 住所変更
住 所 埼玉県川口市本町四丁目1番8号
氏 名 独立行政法人科学技術振興機構